

# Convergence rates of Kernel Conjugate Gradient for random design regression

Gilles Blanchard\*  
 Mathematics Institute, University of Potsdam  
 Karl-Liebknecht-Straße 24-25  
 14476 Potsdam, Germany  
 blanchard@math.uni-potsdam.de

Nicole Krämer  
 Staburo GmbH  
 Aschauer Str. 30a, 81549 München  
 kraemer@staburo.de

## Abstract

We prove statistical rates of convergence for kernel-based least squares regression from i.i.d. data using a conjugate gradient algorithm, where regularization against overfitting is obtained by early stopping. This method is related to Kernel Partial Least Squares, a regression method that combines supervised dimensionality reduction with least squares projection. Following the setting introduced in earlier related literature, we study so-called “fast convergence rates” depending on the regularity of the target regression function (measured by a source condition in terms of the kernel integral operator) and on the effective dimensionality of the data mapped into the kernel space. We obtain upper bounds, essentially matching known minimax lower bounds, for the  $\mathcal{L}^2$  (prediction) norm as well as for the stronger Hilbert norm, if the true regression function belongs to the reproducing kernel Hilbert space. If the latter assumption is not fulfilled, we obtain similar convergence rates for appropriate norms, provided additional unlabeled data are available.

## 1 Introduction

### 1.1 Setting

Consider the nonparametric random design regression (also called “statistical learning”) problem, where an  $n$ -sample of observations  $(X_i, Y_i) \in \mathcal{X} \times \mathbb{R}, 1 \leq i \leq n$  is assumed to be drawn i.i.d. from an unknown distribution  $P$ . Here and in the rest of this work,  $\mathcal{X}$  is assumed to be a Radon space, for instance an open subset of  $\mathbb{R}^d$ . The goal is the estimation of the regression function  $f^*(x) := \mathbb{E}_{(X,Y) \sim P}[Y|X=x]$ ; it is assumed that the true regression function  $f^*$  belongs to the space  $\mathcal{L}^2(\nu)$  of square-integrable functions ( $\nu$  denotes the  $X$ -marginal of  $P$  on the space  $\mathcal{X}$ ).

If  $\hat{f}$  is an estimator of  $f^*$ , its quality is measured via the  $\mathcal{L}^2(\nu)$  distance,

$$\|\hat{f} - f^*\|_{2,\nu}^2 = \mathbb{E}_{X \sim \nu} [(\hat{f}(X) - f^*(X))^2]. \quad (1)$$

---

\*This research was partly supported by the DFG via Research Unit 1735 *Structural Inference in Statistics*.

This distance is natural for random design regression, since, if we interpret this setting as a prediction problem for a new independent example  $(X, Y) \sim P$  where the quality of prediction is measured by the squared error loss  $\ell(f, x, y) = (f(x) - y)^2$ , then it is well-known that  $f^*$  is the minimizer of the average prediction (or generalization) error  $\mathcal{E}(f, P) = \mathbb{E}_{(X, Y) \sim P} [(f(X) - Y)^2]$  over all squared integrable functions, and that the above distance coincides with the excess prediction error:

$$\|\hat{f} - f^*\|_{2, \nu}^2 = \mathcal{E}(\hat{f}) - \mathcal{E}(f^*).$$

Assume that  $k : \mathcal{X}^2 \rightarrow \mathbb{R}$  is a real-valued *reproducing kernel* on the space  $\mathcal{X}$ , with associated reproducing kernel Hilbert space  $\mathcal{H}$ . The well-established principle of non-parametric estimation by reproducing kernel methods consists in considering estimators admitting a *kernel expansion* of the form

$$\hat{f} = f_{\hat{\alpha}}(x) := \frac{1}{n} \sum_{i=1}^n \hat{\alpha}^{(i)} k(X_i, x), \quad (2)$$

where the real coefficients  $\hat{\alpha}^{(i)}$ ,  $1 \leq i \leq n$  are determined from the data (see for example Cristianini and Shawe-Taylor, 2004 and Steinwart and Christmann, 2008 for comprehensive references on the topic.) To avoid some confusion, we point out that the normalization by  $n^{-1}$  that we use here in the kernel expansion is not present in most references on the subject, but we find it technically convenient.

We denote by  $K_n = \frac{1}{n}(k(X_i, X_j))_{i,j} \in \mathbb{R}^{n \times n}$  the normalized kernel matrix and by  $\Upsilon = (Y_1, \dots, Y_n)^\top \in \mathbb{R}^n$  the  $n$ -vector of response observations. A naive approach to determining the vector of kernel expansion coefficients  $\hat{\alpha}$  is to choose those in order that  $f_{\hat{\alpha}}(X_i) = Y_i$  holds for all  $i = 1, \dots, n$ , that is, solving the linear equation

$$K_n \alpha = \Upsilon \quad \text{with } \alpha \in \mathbb{R}^n. \quad (3)$$

Assuming  $K_n$  to be invertible, the solution  $\hat{\alpha}$  of the above equation yields an estimator  $f_{\hat{\alpha}} \in \mathcal{H}$  interpolating perfectly the training data, but that will presumably have very poor performance in terms of the  $\mathcal{L}^2(\nu)$  distance (1), or equivalently having poor generalization error: this is the overfitting phenomenon. There is a variety of possible approaches to counteract this effect by finding a *regularized* solution of (3); perhaps the most well-known one is

$$\hat{\alpha}_\lambda = (K_n + \lambda I)^{-1} \Upsilon, \quad (4)$$

for some fixed parameter  $\lambda > 0$ , known alternatively as kernel ridge regression, Tikhonov's regularization, least squares support vector machine, or MAP Gaussian process regression; for a theoretical study of the convergence rate properties of this approach, see for instance Caponnetto and De Vito (2007), Steinwart and Christmann (2008).

In this paper, we study the conjugate gradient (CG) technique in combination with early stopping to determine the vector of coefficients  $\hat{\alpha}$ . Conjugate gradient is a computationally efficient scheme to approximatively solve linear systems such as (3). The principle of CG is to restrict the problem to a nested set of data-dependent subspaces, the so-called Krylov subspaces, defined as

$$\mathcal{K}_m(\Upsilon, K_n) := \text{vect} \{ \Upsilon, K_n \Upsilon, \dots, K_n^{m-1} \Upsilon \} = \{ p(K_n) \Upsilon, p \in \mathcal{P}_{m-1} \}, \quad (5)$$

where  $\mathcal{P}_{m-1}$  denotes the set of real polynomials of degree at most  $(m-1)$ . Denote by  $\langle \cdot, \cdot \rangle$  the usual euclidean scalar product on  $\mathbb{R}^n$  rescaled by the factor  $n^{-1}$ . We define the  $K_n$ -seminorm as  $\|\alpha\|_{K_n}^2 := \langle \alpha, \alpha \rangle_{K_n} := \langle \alpha, K_n \alpha \rangle$ . Then the CG solution after  $m$  iterations is formally defined as

$$\hat{\alpha}_m = \arg \min_{\alpha \in \mathcal{K}_m(\Upsilon, K_n)} \|\Upsilon - K_n \alpha\|_{K_n}. \quad (6)$$

It is not difficult to prove from (5) and (6) that iterating CG for  $n$  iterations returns  $\hat{\alpha}_n = K_n^\dagger \Upsilon$ , where  $K_n^\dagger$  is the pseudo-inverse of  $K_n$  (in other words, the solution to (3) if  $K_n$  is invertible, and its least squares approximate solution otherwise) and thus will suffer the same overfitting phenomenon mentioned above. However, CG is usually stopped at an early iteration  $m \ll n$ , thus returning an approximate solution. In the learning context considered here, beyond computational aspects making this method attractive, this early stopping is mainly used for its regularization properties. The main contribution of this paper is to study the convergence rates of this approach when the stopping iteration  $m$  is suitably chosen.

Computationally, conjugate gradients have the appealing property that the optimization criterion (6) can be computed by a simple iterative algorithm that constructs basis vectors  $d_1, \dots, d_m$  of  $\mathcal{K}_m(\Upsilon, K_n)$  by using only *forward multiplication* of vectors by the matrix  $K_n$ . Algorithm 1.1 displays the computation of the CG kernel coefficients  $\hat{\alpha}_m$  defined by (6) (see for instance Hanke, 1995, Section 2.2 and Engl et al., 1996, Chapter 7.) Formally, the output of  $m$  steps of the algorithm matches exactly the definition (6). In practice, finite numerical machine precision means that rounding errors can accumulate. Several variations of the algorithm exist, some of which have better reported numerical stability. We will not elaborate more on this topic, since the focus of this paper is on theoretical convergence rates.

---

**Algorithm 1.1:** Kernel Conjugate Gradient regression

---

**Input:** kernel matrix  $K_n$ , response vector  $\Upsilon$ , maximum number of iterations  $m$

**Initialization:**  $\hat{\alpha}_0 = \mathbf{0}_n$ ;  $r_1 = \Upsilon$ ;  $d_1 = \Upsilon$ ;  $t_1 = K_n \Upsilon$ ;

**for**  $i = 1, \dots, m$  **do**

$t_i = t_i / \|t_i\|_{K_n}$ ;  $d_i = d_i / \|t_i\|_{K_n}$  (normalization of the basis, resp. update vector);  
 $\gamma_i = \langle \Upsilon, t_i \rangle_{K_n}$  (proj. of  $\Upsilon$  on basis vector);  
 $\hat{\alpha}_i = \hat{\alpha}_{i-1} + \gamma_i d_i$  (update);  
 $r_{i+1} = r_i - \gamma_i t_i$  (residuals);  
 $d_{i+1} = r_{i+1} - d_i \langle t_i, K_n r_{i+1} \rangle_{K_n}$ ;  $t_{i+1} = K_n d_{i+1}$  (new update, resp. basis vector);

**end**

**Result:** CG kernel coefficients  $\hat{\alpha}_m$ , CG function  $f_m = \sum_{i=1}^n \hat{\alpha}_m^{(i)} k(X_i, \cdot)$

---

## 1.2 Relation to existing work

As we restrict the learning problem onto the Krylov space  $\mathcal{K}_m(\Upsilon, K_n)$ , the CG coefficients  $\hat{\alpha}_m$  are of the form  $\hat{\alpha}_m = q_m(K_n) \Upsilon$  with  $q_m$  a polynomial of degree  $\leq m-1$ . However, the polynomial  $q_m$  is not fixed but depends on  $\Upsilon$  as well, making the CG method nonlinear in the sense that the coefficients  $\hat{\alpha}_m$  depend on  $\Upsilon$  in a nonlinear fashion.

This is in contrast to Tikhonov’s regularization (4), and more generally to the larger family of *spectral linear regularization* methods, which estimate the expansion coefficients via  $\hat{\alpha}_\lambda = F_\lambda(K_n)\Upsilon$ , where  $F_\lambda$  is an appropriately regularized but fixed approximation of the inverse function. For results on the convergence rates of such linear regularization methods in a kernel learning setting comparable to the one studied here, see Bauer et al. (2007); Smale and Zhou (2007); Caponnetto and De Vito (2007); Lo Gerfo et al. (2008); Caponnetto and Yao (2010) and the recent advances Dicker et al. (2015); Blanchard and Mücke (2016). In particular, the convergence rates under source condition type regularity and polynomial eigenvalue decay of the kernel integral operator obtained in the present paper for kernel CG match the rates established for spectral linear regularization methods by Caponnetto and De Vito (2007); Caponnetto and Yao (2010); Dicker et al. (2015) and Blanchard and Mücke (2016).

Both linear regularization methods and the nonlinear CG method are established techniques in the *inverse problem* literature, in a deterministic setting (for a comprehensive overview see Engl et al., 1996.) The statistical kernel learning setting is markedly different since both the design points and the error are stochastic, however the convergence analysis in that setting owes a lot to the mathematical techniques developed in the deterministic case. This is true for linear regularization methods cited above, and holds as well for CG: the present work builds notably on the seminal works of Hanke (1995) and Nemirovskii (1986).

Conjugate gradient methods have appeared under the name of *partial least squares* (PLS) in the statistics literature (Wold et al., 1984), and a “kernelized” version of PLS was developed by Rosipal and Trejo (2001) and is now considered part of the standard toolbox of kernel methods (see Cristianini and Shawe-Taylor, 2004, Section 6.7.2). An important difference with the method we study here is that kernel PLS is defined via (6) but with the  $K_n$ -norm replaced by the regular  $n$ -dimensional Euclidean norm. In conjugate gradient parlance, kernel PLS is “Conjugate Gradient - Minimum Error” (CGME) while we analyze here “Conjugate Gradient applied the the Normal Equations” (CGNE); see Hanke (1995), Section 2.3. Computationally, the two methods are very similar; the main reason we concentrate on kernel CGNE rather than the perhaps more natural kernel CGME is technical: even in the deterministic case, the analysis of CGME presents significantly more technical difficulties (Hanke, 1995, Chap. 4).

The results presented here are an extended version of a prior conference paper (Blanchard and Krämer, 2011). There, a first result was obtained directly based on Nemirovskii’s theorem in the deterministic case: by controlling (via a simple concentration inequality), with high probability, the norm of the errors (on the kernel covariance as well as on the data), it was possible to plug these deterministic estimates directly into Nemirovskii’s theorem, resulting in a bound holding with large probability. However, it was not possible to capture in this way the “fast convergence rate” behavior related to an assumed polynomial decay of the kernel operator’s spectrum, a phenomenon that is specific to the stochastic setting and the object of much attention in the recent years (often under the name “adaptation to the intrinsic data dimensionality”). For reasons of readability, in the present version we decided to skip this first suboptimal (but easy to obtain via Nemirovskii’s theorem) result, to concentrate on the improved fast rate results. The proof of those follows closely the general structure of Nemirovskii’s argument and ideas, but requires a complete reworking in the details due to the additional difficulties arising from taking into account the behavior of the operator’s spectrum. Furthermore, applying Nemirovskii’s result also required to assume  $f^* \in \mathcal{H}$ ,

while the case  $f^* \notin \mathcal{H}$  (called “outer case” below) also introduces additional difficulties. The present version extends the scope of the original conference version by also including convergence results not only in the prediction (or  $\mathcal{L}^2(\nu)$ ) norm, but in stronger norms as well, including the Hilbert  $\mathcal{H}$ -norm when applicable.

## 2 Mathematical framework

### 2.1 Reproducing kernel Hilbert spaces

We assume the reader familiar with the formalism of reproducing kernel Hilbert spaces (RKHS) and refer her for instance to Cristianini and Shawe-Taylor (2004) and Steinwart and Christmann (2008) for more details. We recall briefly a few key points. Given  $k : \mathcal{X}^2 \rightarrow \mathbb{R}$  a real, symmetric and semi-definite positive kernel on  $\mathcal{X}$ , the unique RKHS associated to  $k$  is denoted by  $\mathcal{H}$ . We recall that  $\mathcal{H}$  is a Hilbert space of real-valued functions on  $\mathcal{X}$  containing the functions  $k_x = k(x, \cdot) := (y \in \mathcal{X} \mapsto k(x, y))$  for all  $x \in \mathcal{X}$  and satisfying the characteristic *self-reproducing* property  $\langle k_x, f \rangle_{\mathcal{H}} = f(x)$ , for all  $f \in \mathcal{H}, x \in \mathcal{X}$ . In the rest of this paper we will make the assumption that

**(A)**  $k$  is measurable, for all  $x \in \mathcal{X}$  it holds  $k(x, x) \leq \kappa$ , where  $\kappa$  is a real constant. The Hilbert space  $\mathcal{H}$  is assumed to be separable.

Assumption **(A)** implies that the kernel integral operator

$$K : \mathcal{L}^2(\nu) \rightarrow \mathcal{L}^2(\nu), g \mapsto \int k(\cdot, x)g(x)dP(x), \quad (7)$$

is a well-defined, self-adjoint, Hilbert-Schmidt (and even trace-class) operator. We measure *regularity* of the target function  $f^*$  in terms of a *source condition* with respect to  $K$  and parameters  $\rho > 0, r > 0$ , defined as follows:

$$\mathbf{SC}(r, \rho) : \text{there exists } u \in \mathcal{L}^2(\nu) \text{ such that } f^* = K^r u \quad \text{with} \quad \|u\|_{2, \nu} \leq \kappa^{-r} \rho.$$

Clearly, in the above condition we can assume that  $u \in \text{Ker}(K)^\perp = \overline{\text{Im}(K)}$  without loss of generality. It is well-known that if  $r \geq 1/2$ , then  $f^*$  coincides almost surely with a function belonging to  $\mathcal{H}$ , while for  $r < 1/2$  this is not the case. We refer to  $r \geq 1/2$  as the “inner case” and to  $r < 1/2$  as the “outer case”.

The regularity of the kernel operator  $K$  with respect to the marginal distribution  $\nu$  is measured in terms of the so-called **effective dimensionality** condition. We define the auxiliary notation  $\mathcal{N}(\lambda) := \text{Tr}(K(K + \lambda I)^{-1})$ . Given two parameters  $s \in (0, 1)$ ,  $D \geq 1$ , introduce the condition

$$\mathbf{ED}(s, D) : \mathcal{N}(\lambda) \leq D^2(\kappa^{-1}\lambda)^{-s} \text{ for all } \lambda \in (0, 1].$$

This notion was first introduced by Zhang (2005) in a learning context, and used in a number of works since. It is related to the decay rate of the (ordered) eigenvalues  $(\xi_i)_{i \geq 1}$  of  $K$ : if those satisfy  $\xi_i \leq Ci^{-1/s}$  for some constant  $C$ , then  $\mathbf{ED}(s, D)$  is satisfied for an appropriate constant  $D$ . On the other hand, under the double-sided condition  $ci^{-1/s} \leq \xi_i \leq Ci^{-1/s}$ , lower bounds on

the minimax convergence rates for the model defined by the source conditions **SC**( $r, \rho$ ) are known to be  $\mathcal{O}(n^{-2r/(2r+s)})$  for the  $\mathcal{L}^2(\nu)$  error (Caponnetto and De Vito, 2007), resp.  $\mathcal{O}(n^{-(2r-1)/(2r+s)})$  for the  $\mathcal{H}$ -norm error, assuming  $r \geq 1/2$  (Blanchard and Mücke, 2016).

## 2.2 Conditions on the noise

If  $(X, Y) \sim P$ , denote the noise  $\varepsilon := Y - \mathbb{E}[Y|X]$ . We will consider – depending on the result – one of the following assumptions:

**(Bounded)** (Bounded  $Y$ ):  $|Y| \leq M$  almost surely.

**(Bernstein)** (Bernstein condition):  $\mathbb{E}[\varepsilon^p|X] \leq (1/2)p!M^p$  almost surely, for all integers  $p \geq 2$ .

The second assumption is weaker than the first. In particular, the first assumption implies that not only the noise, but also the target function  $f^*$  is bounded in supremum norm, while the second assumption does not put any additional restriction on the target function.

## 3 Convergence rates

We now introduce the early stopping rule, which takes the form of a so-called **discrepancy stopping rule**: for some threshold  $\Omega > 0$  to be specified, define the (data-dependent) stopping iteration  $\hat{m}$  as the first iteration  $m \geq 0$  (with the convention  $\hat{\alpha}_0 = 0$ ) for which

$$\|\Upsilon - K_n \hat{\alpha}_m\|_{K_n} < \Omega. \quad (8)$$

As mentioned earlier, it holds at the  $n$ -th iteration that  $\hat{\alpha}_n = K_n^\dagger \Upsilon$ , so that  $\|\Upsilon - K_n \hat{\alpha}_n\|_{K_n} = 0$ ; therefore the above stopping rule is well-defined and such that  $\hat{m} \leq n$ . In this section, we assume that the parameters  $r$  and  $s$  appearing in conditions **(SC)** and **(ED)** are known a priori to the user, so that they can be used in the definition of the stopping rule.

Our first result concerns the “inner regularity” case ( $r \geq 1/2$ , so that the target function  $f^*$  coincides a.s. with a function belonging to  $\mathcal{H}$ ),

**Theorem 3.1.** *For some constant  $\tau' > 3/2$  and  $1 > \gamma > 0$ , consider the discrepancy stopping rule with the threshold*

$$\Omega = \tau' M \sqrt{\kappa} \left( \frac{4D}{\sqrt{n}} \log \frac{6}{\gamma} \right)^{\frac{2r+1}{2r+s}}. \quad (9)$$

*Suppose that the noise fulfills the Bernstein assumption **(Bernstein)**, that the source condition **SC**( $r, \rho$ ) holds for  $r \geq 1/2$ , and that **ED**( $s, D$ ) holds. Finally, assume  $n$  is large enough so that*

$$n \geq 16D^2 \log^2(6/\gamma). \quad (10)$$

*Then with probability  $1 - \gamma$ , the estimator  $\hat{f} = f_{\hat{\alpha}_{\hat{m}}}$  obtained by the discrepancy stopping rule (9) satisfies for any  $\theta \in [0, \frac{1}{2}]$ :*

$$\left\| K^{-\theta}(\hat{f} - f^*) \right\|_{2,\nu} \leq c(r, \tau)(\rho + M) \kappa^{-\theta} \left( \frac{4D}{\sqrt{n}} \log \frac{6}{\gamma} \right)^{\frac{2(r-\theta)}{2r+s}}.$$

Observe that in the above result, taking  $\theta = \frac{1}{2}$  results in a convergence rate result in the  $\mathcal{H}$ -norm: this is because for a function  $g$  that coincides a.s. with a function  $g_{\mathcal{H}} \in \mathcal{H}$ , it holds  $\|K^{-\frac{1}{2}}g\|_{2,\nu} = \|g_{\mathcal{H}}\|_{\mathcal{H}}$  (see Section 5.1 for details). Thus we obtain (simultaneous) convergence rate results for the  $\mathcal{L}^2(\nu)$ -norm, the  $\mathcal{H}$ -norm as well as all intermediate norms.

We now turn to the “outer rate” case ( $r < \frac{1}{2}$ ). In this situation, following an idea used by Caponnetto and Yao (2010), we make the additional assumption that *unlabeled* data is available. Assume that we have  $\tilde{n}$  i.i.d. observations  $X_1, \dots, X_{\tilde{n}}$ , out of which only the first  $n$  are labeled. We define a new response vector  $\bar{\mathbf{Y}} = \frac{\tilde{n}}{n} (Y_1, \dots, Y_n, 0, \dots, 0) \in \mathbb{R}^{\tilde{n}}$  and run the CG algorithm 1.1 on  $X_1, \dots, X_{\tilde{n}}$  and  $\bar{\mathbf{Y}}$ . We use the stopping rule with the following threshold:

$$\Omega = \tau' \max(\rho, M) \sqrt{\kappa} \left( \frac{4D}{\sqrt{n}} \log \frac{6}{\gamma} \right)^{\frac{2r+1}{2r+s}}. \quad (11)$$

Observe that it is similar to (9) in the previous section, except the factor  $M$  is replaced by  $\max(M, \rho)$  (and the numerical constants are different).

**Theorem 3.2.** *For some constant  $\tau' > 6$ , and  $\gamma \in (0, 1)$ , consider the discrepancy stopping rule with the fixed threshold given by (11).*

*Suppose assumptions **(Bounded)**, **SC**( $r, \rho$ ) and **ED**( $s, D$ ), are granted with  $r < \frac{1}{2}$  and  $r + s \geq \frac{1}{2}$ . Assume  $n$  is large enough so that*

$$n \geq 16D^2 \log^2(4/\gamma), \quad (12)$$

*and that additional unlabeled data is available with  $\tilde{n} \geq n^{\frac{1+s}{2r+s}}$ . Then with probability  $1 - \gamma$ , the estimator  $\hat{f}$  obtained by the discrepancy stopping rule defined above satisfies for any  $\theta \in [0, r)$ :*

$$\left\| K^{-\theta}(\hat{f} - f^*) \right\|_{2,\nu} \leq c(r, \tau)(\rho + M) \kappa^{-\theta} \left( \frac{4D}{\sqrt{n}} \log \frac{4}{\gamma} \right)^{\frac{2(r-\theta)}{2r+s}}.$$

In the outer case, since  $f^* \notin \mathcal{H}$  we of course cannot expect any convergence in  $\mathcal{H}$ -norm, but as is clear from the above result, we obtain convergence rate results in norms that are stronger than the  $\mathcal{L}^2(\nu)$ -norm, with the meaningful range of  $\theta$  (determining the strength of the norm) determined by the source condition parameter  $r$ .

## 4 Discussion

*Rate quasi-optimality.* Convergence rates are generally stated in expectation, while the convergence results of Theorems 3.1 and 3.2 are stated with high probability. Usually, exponential deviation bounds can be integrated to yield bounds in expectation; unfortunately, this is not directly possible here, because (a) conditions (10) (resp. (12)) introduce a constraint between the number of examples  $n$  and the probability  $\gamma$  that the bound fails, preventing a statement about extreme (exponentially small in  $n$ ) quantiles of the error; and even more importantly (b) the threshold  $\Omega$  for the stopping criterion itself depends on the prescribed bound failure probability  $\gamma$ . For the present discussion, we therefore consider the following slightly weaker notion considered by Caponnetto and



De Vito (2007): we call a positive sequence  $(a_{n,\theta})_{n \geq 1}$  an upper rate of convergence in probability for the sequence of estimators  $\hat{f}_n$  over a class of distributions  $\mathcal{P}$  if

$$\lim_{C \rightarrow \infty} \limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}} \mathbb{P}_{(X_i, Y_i)_{i=1}^n \stackrel{i.i.d.}{\sim} P} \left[ \left\| K^{-\theta}(\hat{f}_n - f_P^*) \right\|_{2,\nu} > C a_{n,\theta} \right] = 0, \quad (13)$$

On the other hand, for the class of distributions defined by the source condition  $\mathbf{SC}(r, \rho)$  and the polynomial decay condition  $ci^{-1/s} \leq \xi_i \leq Ci^{-1/s}$  for the eigenvalues  $(\xi_i)_{i \geq 1}$  of the kernel integral operator  $K$ , the convergence rate  $a_{n,\theta}^* := n^{-\frac{r-\theta}{2r+s}}$  is minimax optimal (for lower bounds on attainable rates see, for instance, Caponnetto and De Vito, 2007 for the case  $\theta = 0$  and Blanchard and Mücke, 2016 for the case  $r \geq \frac{1}{2}, \theta \in [0, \frac{1}{2}]$ .)

We can thus conclude that kernel CG enjoys quasi-optimal statistical rates of convergence, in the sense that any sequence  $a_{n,\theta}$  such that  $a_{n,\theta}^* = o(a_{n,\theta})$  can be an upper rate of convergence, provided the stopping iteration is chosen appropriately. Namely, we can choose the sequence of bound failure probabilities  $(\gamma_n)_{n \geq 1}$  converging to 0 arbitrarily slowly, so that the rate obtained using the corresponding sequence  $\Omega_n$  in Theorems 3.1 or 3.2 yield (13) with a rate  $a_{n,\theta} = a_{n,\theta}^* c_n$  where  $c_n$  tends arbitrarily slowly to  $\infty$ .

*Comparison of methods.* It is known from previous works that a large family of spectral linearization methods (see in particular Caponnetto and De Vito, 2007 for Tikhonov regularization, and Caponnetto and Yao, 2010; Blanchard and Mücke, 2016 for the general case) achieve minimax optimal convergence rates in the setting considered here. It is therefore a legitimate question whether the additional technicality for analyzing CG is justified, given the availability of other methods. The main reason is that CG remains an algorithm of choice because of its excellent computational properties. Because it very aggressively aims at reducing the residual error, it is often observed in practice that CG converges in much fewer iterations than other methods (such as regular gradient descent, studied in the kernel learning setting by Yao et al., 2007). For this reason it would be of interest to analyze the convergence rate of kernel PLS as well, which, as already mentioned in the introduction, is computationally very similar to kernel CG but more challenging to study theoretically.

*Adaptivity.* The quasi-optimal convergence rates obtained in this work use a stopping criterion with a threshold depending on the parameters  $r$  and  $s$  (and on  $\rho$  in the outer case). It is unrealistic to assume these various regularity parameters to be known in advance in practice. The question of automatic choice of a stopping iteration without prior knowledge of these parameters is known as that of adaptivity. For what concerns the convergence in excess prediction error  $\mathcal{L}^2(\nu)$ , and under the assumption **(Bounded)**:  $|Y| \leq M$ , it is well known that a simple hold-out strategy (i.e. choosing, amongst a family of candidate estimators, the one achieving minimal error on an held-out validation sample), performed after trimming all candidate estimators  $\hat{f}_m$  to the interval  $[-M, M]$ , generally speaking selects an estimator close to the best between those considered. In the present context, each iteration  $m$  provides such a candidate estimate; one could for example adapt the corresponding arguments from Caponnetto and Yao (2010); see also Blanchard and Massart (2006) for a general point of view on this question. As a consequence, since the results established here grant the *existence* of an iteration with quasi-optimal rate, this will also be the case for the adaptive hold-out strategy. It remains an open question whether this also applies to



the error measured in stronger norms: although we established that there exists an iteration with optimal rates for all norms, it does not follow that any iteration which is good in the sense of excess prediction (one of which hold-out would select) would automatically also yield good performance for the stronger norms.

## 5 Proofs

The proof of our results rely on combining ideas of Hanke (1995) (expanding upon the remarkable seminal work of Nemirovskii, 1986), used in the analysis of convergence of CG algorithms for (deterministic) inverse problems, with tools introduced by De Vito et al. (2006), Caponnetto and De Vito (2007), Smale and Zhou (2007), Caponnetto and Yao (2010) for the analysis of inverse problem methods for the statistical learning setup. We start by gathering in the two next sections the required notation and previous results that we will make use of, before getting to the proof itself. In all the proofs, we use the notation  $c(a, b)$  to denote a nonrandom function only depending on the nonrandom parameters  $a, b$ , and whose exact value can change from line to line.

### 5.1 Setup and key tools for statistical learning as an inverse problem

We first define the *empirical evaluation operator*  $T_n$  as follows:

$$T_n : \quad g \in \mathcal{H} \mapsto T_n g := (g(X_1), \dots, g(X_n))^{\top} \in \mathbb{R}^n$$

and the *empirical integral operator*  $T_n^*$  as:

$$T_n^* : u = (u_1, \dots, u_n) \in \mathbb{R}^n \mapsto T_n^* u := \frac{1}{n} \sum_{i=1}^n u_i k(X_i, \cdot) \in \mathcal{H}.$$

Using the reproducing property of the kernel, it can be readily checked that  $T_n$  and  $T_n^*$  are adjoint operators, i.e. they satisfy  $\langle T_n^* u, g \rangle_{\mathcal{H}} = \langle u, T_n g \rangle$ , for all  $u \in \mathbb{R}^n, g \in \mathcal{H}$ . With this notation, it is clear that if  $\hat{\alpha} \in \mathbb{R}^n$  is the vector of coefficients in the normalized kernel expansion (2) of a kernel estimator  $\hat{f}$ , then it holds  $\hat{f} = T_n^* \hat{\alpha}$ . Furthermore, since  $K_n = T_n T_n^*$ , we have for any  $u \in \mathbb{R}^n$ :

$$\|u\|_{K_n}^2 = \langle u, K_n u \rangle = \|T_n^* u\|_{\mathcal{H}}^2.$$

Based on these facts, equation (6) can be rewritten as

$$\hat{\alpha}_m = \arg \min_{\alpha \in \mathcal{K}_m(\Upsilon, K_n)} \|T_n^* \Upsilon - T_n^* T_n T_n^* \alpha\|_{\mathcal{H}},$$

implying that for the  $m$ -th iteration estimator  $f_m = T_n^* \hat{\alpha}_m$ , it holds

$$f_m = \arg \min_{f \in \mathcal{K}_m(T_n^* \Upsilon, S_n)} \|T_n^* \Upsilon - S_n f\|_{\mathcal{H}}, \quad (14)$$

where  $S_n = T_n^* T_n$  is a self-adjoint operator of  $\mathcal{H}$ , called empirical covariance operator. In the sequel we will mainly refer to (14) as the characterization of the CG method. In fact, (14) corresponds to

the definition of the “usual” conjugate gradient algorithm (in Hilbert space), formally applied to the so-called normal equation (in  $\mathcal{H}$ )

$$S_n \hat{f} = T_n^* \Upsilon, \quad (15)$$

which is obtained from (3) by left multiplication by  $T_n^*$ .

The advantage of this reformulation, and an idea first introduced by De Vito et al. (2006), is that it can be interpreted as a perturbation of a *population, noiseless* version (of the equation and of the algorithm), wherein  $\Upsilon$  is replaced by the target function  $f^*$  and the empirical operators  $T_n^*, T_n$  are respectively replaced by their population analogues, the kernel integral operator

$$T^* : g \in \mathcal{L}^2(\nu) \mapsto T^* g := \int k(x, \cdot) g(x) d\nu(x) = \mathbb{E} [k(X, \cdot) g(X)] \in \mathcal{H},$$

and the change-of-space (or inclusion) operator

$$T : g \in \mathcal{H} \mapsto g \in \mathcal{L}^2(\nu).$$

The latter maps a function to itself but between two Hilbert spaces which differ with respect to their geometry – the inner product of  $\mathcal{H}$  being defined by the kernel function  $k$ , while the inner product of  $\mathcal{L}^2(\nu)$  depends on the data generating distribution. This operator is well defined: since the kernel is bounded, all functions in  $\mathcal{H}$  are bounded and therefore square integrable under any distribution  $\nu$ ; this also implies that  $T^*$  is well-defined. Again, it can be checked due to the reproducing property that  $T, T^*$  are adjoint of each other; we denote  $S := T^* T$  the population covariance operator, and observe that  $K = T T^*$  holds, where  $K$  is the operator defined by (7). Finally, it holds that  $S^{-\frac{1}{2}} T^*$  is a partial isometry from  $\mathcal{L}^2(\nu)$  to  $\mathcal{H}$ , and  $K^{-\frac{1}{2}} T$  a partial isometry from  $\mathcal{H}$  to  $\mathcal{L}^2(\nu)$ . In particular, if  $g \in \mathcal{L}^2(\nu)$  coincides a.s. with some function  $g_{\mathcal{H}} \in \mathcal{H}$ , then it holds  $g = T g_{\mathcal{H}}$  with  $\|g_{\mathcal{H}}\|_{\mathcal{H}} = \|K^{-\frac{1}{2}} T g_{\mathcal{H}}\|_{2, \nu} = \|K^{-\frac{1}{2}} g\|_{2, \nu}$ .

The next lemma was established by Caponnetto and De Vito (2007), based on a Bernstein-type inequality for random variables taking values in a Hilbert space, see Pinelis and Sakhanenko (1985); Yurinski (1995). It bounds with high probability the deviations between the quantities in the normal equations (15) and their population counterparts. A key insight from Caponnetto and De Vito (2007) is that in order to obtain sharp bounds on convergence rates, these deviations should be measured in a “warped” norm rather than in the standard norm:

**Lemma 5.1.** *Let  $\lambda$  be a positive number. Under assumption **(Bounded)**, the following holds:*

$$\mathbb{P} \left[ \left\| (S + \lambda I)^{-\frac{1}{2}} (T_n^* \Upsilon - T^* f^*) \right\|_{\mathcal{H}} \leq 2M \left( \sqrt{\frac{\mathcal{N}(\lambda)}{n}} + \frac{2\sqrt{\kappa}}{\sqrt{\lambda n}} \right) \log \frac{2}{\gamma} \right] \geq 1 - \gamma. \quad (16)$$

*If the representation  $f^* = T f_{\mathcal{H}}^*$  holds and under assumption **(Bernstein)**, we have the following:*

$$\mathbb{P} \left[ \left\| (S + \lambda I)^{-\frac{1}{2}} (T_n^* \Upsilon - S_n f_{\mathcal{H}}^*) \right\|_{\mathcal{H}} \leq 2M \left( \sqrt{\frac{\mathcal{N}(\lambda)}{n}} + \frac{2\sqrt{\kappa}}{\sqrt{\lambda n}} \right) \log \frac{2}{\gamma} \right] \geq 1 - \gamma. \quad (17)$$

Concerning the convergence of empirical covariance in the sense of operators, the following holds:

$$\mathbb{P} \left[ \left\| (S + \lambda I)^{-\frac{1}{2}} (S_n - S) \right\|_{HS} \leq 2\sqrt{\kappa} \left( \sqrt{\frac{\mathcal{N}(\lambda)}{n}} + \frac{2\sqrt{\kappa}}{\sqrt{\lambda n}} \right) \log \frac{2}{\gamma} \right] \geq 1 - \gamma, \quad (18)$$

as well as:

$$\mathbb{P} \left[ \|S_n - S\|_{HS} \leq \frac{4\kappa}{\sqrt{n}} \sqrt{\log \frac{2}{\gamma}} \right] \geq 1 - \gamma, \quad (19)$$

where  $\|\cdot\|_{HS}$  denotes the Hilbert-Schmidt norm.

## 5.2 Key tools for the analysis of CG using orthogonal polynomials theory

We denote by  $(\xi_{n,i}, e_{n,i})_{i \in I}$  (respectively  $(\xi_i, e_i)_{i \in I}$ ) an eigenvalue-eigenvector orthogonal basis with  $\xi_{n,i}, \xi_i \in [0, \kappa]$  for the operator  $S_n$ , respectively  $S$ . Since the rank of  $S_n$  is at most  $n$ , the family  $(\xi_{n,i})_{i \in I}$  has at most  $n$  nonzero terms (while the family  $(\xi_i)_{i \in I}$  has at most countably many nonzero terms,  $S$  being compact).

Using the formalism of functional calculus for operators, if  $\phi : [0, \kappa] \rightarrow \mathbb{R}$  is a bounded and measurable function, we denote

$$\phi(S_n) := \sum_{i \in I} \phi(\xi_{n,i}) e_{n,i} e_{n,i}^* \quad \text{and} \quad \phi(S) := \sum_{i \in I} \phi(\xi_i) e_i e_i^*.$$

We recall the “switching” rule  $T\phi(S) = \phi(K)T^*$  since  $S = T^*T$ ,  $K = TT^*$ , which we will be using often. In the sequel,  $\|A\|$  denotes the operator norm of an operator  $A$ . The above definition implies in particular the bound

$$\|\phi(S_n)\| \leq \sup_{t \in [0, \kappa]} |\phi(t)|.$$

For  $u \geq 0$ , denote  $F_u = \mathbf{1}_{[0, u)}(S_n)$ , that is, the orthogonal projector in  $\mathcal{H}$  onto the subspace  $\text{vect}\{e_{n,i}, i \in I \text{ s.t. } \xi_{n,i} < u\}$  spanned by eigenvectors of  $S_n$  corresponding to eigenvalues strictly less than  $u$ . Observe that  $F_u$  depends on the data because  $S_n$  does, but we omit the index  $n$  at this juncture to simplify notation, and because we will not make use of the corresponding notion for  $S$ , so that there is no risk of confusion. Finally, for an integer  $\ell$  introduce the measure

$$\mu_n^{(\ell)} := \sum_{i \in I} \xi_{n,i}^\ell \langle T_n^* \Upsilon, e_{n,i} \rangle^2 \delta_{\xi_{n,i}},$$

where  $\delta_x$  denotes the Dirac delta-measure at point  $x$ . In particular, for  $\ell = 0$  we use the convention  $0^0 = 1$  and it holds for a bounded measurable function  $\phi : [0, \kappa] \rightarrow \mathbb{R}$ :

$$\|\phi(S_n) T_n^* \Upsilon\|_{\mathcal{H}}^2 = \sum_{i \in I} \phi(\xi_{n,i})^2 \langle T_n^* \Upsilon, e_{n,i} \rangle^2 = \int \phi(t)^2 d\mu_n^{(0)}(t).$$

Observe that, for  $i \in I$  such that  $\xi_{n,i} = 0$ , we have  $\langle T_n^* \Upsilon, e_{n,i} \rangle = \langle \Upsilon, T_n e_{n,i} \rangle = 0$  since  $e_{n,i} \in \ker(T_n^* T_n) = \ker(T_n)$ . Therefore, the measures  $\mu_n^{(\ell)}$  have finite support (independent of  $\ell$ ) of cardinality  $n_\Upsilon \leq n$ . In fact  $n_\Upsilon$  is the number of distinct positive eigenvalues of  $K_n$  such that  $\Upsilon$

has nonzero projection on the corresponding eigenspace (or equivalently, the number of distinct positive eigenvalues of  $S_n$  such that  $T_n^* \Upsilon$  has nonzero projection on the corresponding eigenspace).

With this formalism established, we turn to properties of the CG method (see, e.g., Hanke, 1995, and Engl et al., 1996, Chapter 7). By its definition, the output of the  $m$ -th iteration of the CG algorithm can be put under the form  $f_m = q_m(S_n)T_n^* \Upsilon$ , where  $q_m \in \mathcal{P}_{m-1}$ , the vector space of real polynomials of degree at most  $m-1$ . A crucial role is played by the *residual polynomial*

$$p_m(x) = 1 - xq_m(x) \in \mathcal{P}_m^0,$$

where  $\mathcal{P}_m^0$  is the affine space of real polynomials of degree no greater than  $m$  and having constant term equal to 1. In particular  $T_n^* \Upsilon - S_n f_m = p_m(S_n)T_n^* \Upsilon$ . For  $\ell \geq 0$  we define

$$\begin{aligned} [p, q]_{(\ell)} &:= \int_0^\kappa p(t)q(t)d\mu_n^{(\ell)}(t) = \left\langle p(S_n)T_n^* \Upsilon, S_n^\ell q(S_n)T_n^* \Upsilon \right\rangle \\ &= \sum_{i \geq 1} p(\xi_{n,i})q(\xi_{n,i})\xi_{n,i}^\ell \langle T_n^* \Upsilon, e_{n,i} \rangle^2. \end{aligned}$$

Since the measure  $\mu_n^{(\ell)}$  has support of cardinality  $n_\Upsilon$ ,  $[\cdot, \cdot]_{(\ell)}$  is a scalar product on the space  $\mathcal{P}_{n_\Upsilon-1}$ . Consider an iteration  $m < n_\Upsilon$ . By (14),  $q_m$  is the minimizer of  $\|(I - S_n q(S_n))T_n^* \Upsilon\|_{\mathcal{H}}^2$  over  $q \in \mathcal{P}_{m-1}$ , so that the residual polynomial  $p_m$  is equivalently a minimizer of  $\|p(S_n)T_n^* \Upsilon\|_{\mathcal{H}}^2 = [p, p]_{(0)}$  over  $p \in \mathcal{P}_m^0$ . In other words,  $p_m$  is the orthogonal projection of the origin onto the affine subspace  $\mathcal{P}_m^0 \subset \mathcal{P}_m$  for the scalar product  $[\cdot, \cdot]_{(0)}$ . This, in passing, shows the unicity of  $p_m$ , and by consequence of  $q_m$  and  $f_m$ . In the case  $m = 0$ , we set  $q_0 = 0, p_0 \equiv 1$ .

We denote by  $\pi$  the shift operation on polynomials with  $(\pi q)(x) = xq(x)$ . Since  $\mathcal{P}_m^0 = 1 + \pi\mathcal{P}_{m-1}$  is an affine subspace of  $\mathcal{P}_m$  parallel to  $\pi\mathcal{P}_{m-1}$ , it follows by the properties of projections that  $p_m$  is orthogonal to  $\pi\mathcal{P}_{m-1}$  for  $[\cdot, \cdot]_{(0)}$ . Thus  $0 = [p_m, \pi q]_{(0)} = [p_m, q]_{(1)}$  for any  $q \in \mathcal{P}_{m-1}$ ; this establishes that  $p_0, p_1, \dots, p_{n_\Upsilon-1}$  is an orthogonal polynomial sequence with respect to  $[\cdot, \cdot]_{(1)}$ . For  $m = n_\Upsilon$ , this is somewhat of a special case since  $[\cdot, \cdot]_{(\ell)}$  is only a semidefinite product on  $\mathcal{P}_{n_\Upsilon}$ . However, it is not difficult to see that the polynomial  $p_{n_\Upsilon}$  having  $n_\Upsilon$  distinct roots corresponding to the atoms of  $\mu_n^{(0)}$  and normalized to have constant term equal to 1, is the unique element of  $\mathcal{P}_{n_\Upsilon}^0$  satisfying  $[p_{n_\Upsilon}, p_{n_\Upsilon}]_{(0)} = 0$ . Therefore, unicity of the solution also holds for  $m = n_\Upsilon$ , and obviously also  $[p_{n_\Upsilon}, p_m]_{(1)} = 0$  for all  $m \leq n_\Upsilon$ . The CG method will not in any case go beyond iteration  $m = n_\Upsilon$ , since at this point, by the above considerations the residual norm is 0 and an exact solution to the equation  $S_n f = T_n^* \Upsilon$  has been reached.

The next lemma gathers the technical results coming from the theory of orthogonal polynomials needed for our analysis.

**Lemma 5.2.** *Let  $m$  be any integer satisfying  $1 \leq m \leq n_\Upsilon$ .*

- i) The polynomial  $p_m$  has exactly  $m$  distinct roots belonging to  $(0, \kappa]$ , denoted by  $(x_{k,m})_{1 \leq k \leq m}$  in increasing order.*
- ii)  $p_m$  is positive, decreasing and convex on the interval  $[0, x_{1,m})$ .*
- iii) Define the function  $\varphi_m$  on the interval  $[0, x_{1,m})$  as*

$$\varphi_m(x) = p_m(x) \left( \frac{x_{1,m}}{x_{1,m} - x} \right)^{\frac{1}{2}}.$$

Then it holds

$$[p_m, p_m]_{(0)}^{\frac{1}{2}} = \|p_m(S_n)T_n^*\Upsilon\|_{\mathcal{H}} \leq \|F_{x_{1,m}}\varphi_m(S_n)T_n^*\Upsilon\|_{\mathcal{H}}, \quad (20)$$

and furthermore, for any  $\nu \geq 0$  (and the convention  $0^0 = 1$ ):

$$\sup_{x \in [0, x_{1,m}]} x^\nu \varphi_m^2(x) \leq \nu^\nu |p'_m(0)|^{-\nu}. \quad (21)$$

iv) Denote  $p_0^{(2)}, p_1^{(2)}, \dots, p_{n_\Upsilon-1}^{(2)}$  the unique sequence of orthogonal polynomials with respect to  $[\cdot, \cdot]_{(2)}$  and with constant term equal to 1. This sequence enjoys properties (i) and (ii) above, with  $(x_{k,m}^{(2)})_{1 \leq k \leq m}$  denoting the distinct roots of  $p_m^{(2)}$  in increasing order. Then it holds that  $x_{1,m} \leq x_{1,m}^{(2)}$ . Finally, the following holds (Christoffel-Darboux identity):

$$0 \leq p'_{m-1}(0) - p'_m(0) = \frac{[p_{m-1}, p_{m-1}]_{(0)} - [p_m, p_m]_{(0)}}{\left[p_{m-1}^{(2)}, p_{m-1}^{(2)}\right]_{(1)}} \leq \frac{[p_{m-1}, p_{m-1}]_{(0)}}{\left[p_{m-1}^{(2)}, p_{m-1}^{(2)}\right]_{(1)}}. \quad (22)$$

For a proof of these properties see the monograph of Hanke (1995), from which the above properties have been collected. Existence of a unique family of orthogonal polynomials  $p_k^{(\ell)}$  for any  $\ell \geq 0$ , up to degree  $n_\Upsilon - 1$ , is guaranteed by the fact that the measures  $\mu_n^{(\ell)}$  have support of cardinality  $n_\Upsilon$ . Point (i) is well-known in the theory of orthogonal polynomials, see also Hanke (1995), Section 2.4. Point (ii) is equally well-known and an easy consequence of (i), namely (ii) holds true for any real polynomial of degree  $m$  having  $m$  real positive roots and taking a positive value at 0, due to the interlacing property of the roots of its derivatives. For point (iv), all roots of  $p_m^{(2)}$  are positive by standard results of orthogonal polynomial theory, so we can normalize these polynomials to have constant term equal to 1. Relation (22), resp. the relation  $x_{1,m} \leq x_{1,m}^{(2)}$  can be found as Corollary 2.6, resp. 2.7, of Hanke (1995). Finally, point (iii) can be found as an ingredient of the proof of Lemma 3.7 of Hanke (1995), more precisely (20),(21) are found respectively as (3.8) and (3.10) there (or equivalently as (7.7) and (7.8) in Chapter 7 of Engl et al., 1996). The seminal idea of introducing the function  $\varphi_m$  above and properties (20)-(21) are originally due to Nemirovskii (1986).

### 5.3 Proof of Theorem 3.1

We recall that since we assume  $r \geq 1/2$ , there exists  $f_{\mathcal{H}}^* \in \mathcal{H}$  such that  $f^* = Tf_{\mathcal{H}}^*$  holds. The main effort below is to analyze the algorithm when the events of high probability of Lemma 5.1 are satisfied. To simplify notation, we will define the following event, where  $\Lambda \geq 1, \Delta \geq 0$  are constants and  $\delta(\lambda) \geq 0$  only depends on  $\lambda$ :

$$\mathbf{B}(\lambda) : \begin{cases} \left\| (S + \lambda I)^{-\frac{1}{2}} (T_n^* \Upsilon - S_n f_{\mathcal{H}}^*) \right\|_{\mathcal{H}} & \leq \delta(\lambda), \\ \left\| (S + \lambda I)(S_n + \lambda I)^{-1} \right\|_{HS} & \leq \Lambda^2, \\ \|S - S_n\|_{HS} & \leq \kappa \Delta. \end{cases}$$

In the rest of this proof we set  $\mu = r - 1/2$ . Under the source condition assumption  $\mathbf{SC}(r, \rho)$ , for  $r \geq \frac{1}{2}$  the representation  $f^* = K^r u$  can be rewritten

$$f^* = (TT^*)^r u = T(T^*T)^{r-\frac{1}{2}}(T^*T)^{-\frac{1}{2}}T^*u = TS^\mu S^{-\frac{1}{2}}T^*u,$$

by identification we therefore have the source condition for  $f_{\mathcal{H}}$  given by  $f_{\mathcal{H}} = S^\mu w$  with  $w = S^{-\frac{1}{2}}T^*u$ , and  $\|w\|_{\mathcal{H}} \leq \|u\|_{2,\nu} \leq \kappa^{-\mu-\frac{1}{2}}\rho$ , since  $S^{-\frac{1}{2}}T^*$  is a partial isometry from  $\mathcal{L}^2(\nu)$  into  $\mathcal{H}$ .

Finally we define following shortcut notation for  $\beta > 0$ :

$$Z_\beta(\lambda) = \begin{cases} \lambda^\beta & \text{for } \beta \leq 1, \\ \kappa^\beta \Delta & \text{for } \beta > 1. \end{cases} \quad (23)$$

In order to simplify notation, for the remainder of the paper we will omit the indices from  $\|\cdot\|_{\mathcal{H}}$  and  $\|\cdot\|_{2,\nu}$ ; in other words the notation  $\|\cdot\|$  will be overloaded to mean operator,  $\mathcal{H}$  or  $\mathcal{L}^2(\nu)$  norm, depending (nonambiguously) on the context. Note that the  $\mathcal{L}^2(\nu)$  norm will not be explicitly used again until the proof of Theorem 3.2 in Section 5.4.

We start with a technical lemma encapsulating a couple of bounding devices that we will use repeatedly.

**Lemma 5.3.** *Let  $\lambda > 0$  be fixed. Assume the event  $\mathbf{B}(\lambda)$  is satisfied. For any  $\nu \in [0, 1]$ , it holds*

$$\|(S + \lambda I)^\nu (S_n + \lambda I)^{-\nu}\| \leq \Lambda^{2\nu}. \quad (24)$$

*For any  $\nu \in [0, 1]$ , and any  $h \in \mathcal{H}$ , it holds*

$$\|S^\nu h\| \leq \Lambda^{2\nu} \|(S_n + \lambda)^\nu h\|. \quad (25)$$

*For any  $\nu > 0$  and for any  $\phi : [0, \kappa] \rightarrow \mathbb{R}$  measurable function, it holds*

$$\|\phi(S_n)S^\nu\| \leq \Lambda^2 \left( \sup_{t \in [0, \kappa]} t^\nu \phi(t) + (\nu \vee 1) Z_\nu(\lambda) \sup_{t \in [0, \kappa]} \phi(t) \right). \quad (26)$$

*Proof.* Inequality (24) is a direct consequence of the second component in event  $\mathbf{B}(\lambda)$ , and of the operator norm inequality  $\|A^\nu B^\nu\| \leq \|AB\|^\nu$  for self-adjoint positive operators. See Bathia (1997), Theorem X.1.1, where the result is stated for positive matrices, but the proof applies as well to positive operators on a Hilbert space. For the second inequality, we have

$$\|S^\nu h\| \leq \|S^\nu (S + \lambda I)^{-\nu}\| \|(S + \lambda I)^\nu (S_n + \lambda I)^{-\nu}\| \|(S_n + \lambda I)^\nu h\| \leq \Lambda^{2\nu} \|(S_n + \lambda)^\nu h\|.$$

Concerning the last part of the lemma, we first consider the case  $\nu > 1$ ; then

$$\begin{aligned} \|\phi(S_n)S^\nu\| &\leq \|\phi(S_n)\| \|S^\nu - S_n^\nu\| + \|\phi(S_n)S_n^\nu\| \\ &\leq \sup_{t \in [0, \kappa]} t^\nu \phi(t) + \|(S^\nu - S_n^\nu)\| \sup_{t \in [0, \kappa]} \phi(t) \end{aligned}$$

Furthermore, we have

$$\|(S^\nu - S_n^\nu)\| \leq \|(S^\nu - S_n^\nu)\|_{HS} \leq \nu \kappa^{\nu-1} \|S - S_n\|_{HS} \leq \nu \kappa^\nu \Delta.$$

The second inequality used that if  $A, B$  are two semipositive self-adjoint Hilbert-Schmidt operators, and  $\phi$  is a  $L$ -Lipschitz function on  $[0, \max(\|A\|, \|B\|)]$ , then  $\|\phi(A) - \phi(B)\|_{HS} \leq L \|A - B\|_{HS}$  (see, for instance, Bathia, 1997, Lemma VII.5.5, for a proof in the finite dimensional case that can be easily extended to the Hilbert-Schmidt case. Note in passing that this inequality does not hold for the operator norm in general). We applied this property to the power function  $x \mapsto x^\nu$ , which is  $\nu \kappa^{\nu-1}$ -Lipschitz over  $[0, \kappa]$ .

In the case  $\nu \leq 1$ , we have

$$\begin{aligned} \|\phi(S_n)S^\nu\| &\leq \|\phi(S_n)(S_n + \lambda I)^\nu\| \|(S_n + \lambda I)^{-\nu}(S + \lambda I)^\nu\| \|(S + \lambda I)^{-\nu}S^\nu\| \\ &\leq \Lambda^{2\nu} \left( \sup_{t \in [0, \kappa]} t^\nu \phi(t) + \lambda^\nu \sup_{t \in [0, \kappa]} \phi(t) \right). \end{aligned}$$

□

**Lemma 5.4** (Bounding the error). *Assume condition  $\mathbf{SC}(r, \rho)$  holds,  $r \geq \frac{1}{2}$ . For any  $\lambda > 0$ , if the event  $\mathbf{B}(\lambda)$  is satisfied, then for any iteration step  $1 \leq m \leq n_\Upsilon$  and  $\theta \in [0, \frac{1}{2}]$ , for any  $\varepsilon \in (0, x_{1,m})$ , and denoting  $\tilde{\varepsilon} := \min(\varepsilon, |p'_m(0)|^{-1})$ :*

$$\begin{aligned} \|S^{\frac{1}{2}-\theta}(f_m - f_{\mathcal{H}}^*)\| &\leq c(\Lambda, \mu) \left( \tilde{\varepsilon}^{-1} (\tilde{\varepsilon} + \lambda)^{1-\theta} \delta(\lambda) + (\varepsilon^\mu + Z_\mu(\lambda)) (\varepsilon + \lambda)^{\frac{1}{2}-\theta} \kappa^{-\mu-\frac{1}{2}} \rho \right. \\ &\quad \left. + \varepsilon^{-1} (\varepsilon + \lambda)^{\frac{1}{2}-\theta} \|T_n^*(T_n f_m - \Upsilon)\| \right) \end{aligned}$$

For  $m = 0$ , the above inequality is valid for any  $\varepsilon > 0$ .

*Proof.* Set  $\bar{f}_m = q_m(S_n)S_n f_{\mathcal{H}}^*$ . This is the element in  $\mathcal{H}$  that we obtain by applying the  $m$ th-iteration CG polynomial  $q_m$  to the *noiseless* data. We have using (25)

$$\begin{aligned} \|S^{\frac{1}{2}-\theta}(f_m - f_{\mathcal{H}}^*)\| &\leq \Lambda^{1-2\theta} \|(S_n + \lambda I)^{\frac{1}{2}-\theta}(f_m - f_{\mathcal{H}}^*)\| \\ &\leq \Lambda \left( \|F_\varepsilon(S_n + \lambda I)^{\frac{1}{2}-\theta}(f_m - \bar{f}_m)\| + \|F_\varepsilon(S_n + \lambda I)^{\frac{1}{2}-\theta}(\bar{f}_m - f_{\mathcal{H}}^*)\| \right. \\ &\quad \left. + \|F_\varepsilon^\perp(S_n + \lambda I)^{\frac{1}{2}-\theta}(f_m - f_{\mathcal{H}}^*)\| \right) \\ &:= \Lambda((I) + (II) + (III)), \end{aligned}$$

where we denote  $F_\varepsilon^\perp := (I - F_\varepsilon)$ . We upper bound the first summand and start with using the



first component of event  $\mathbf{B}(\lambda)$ :

$$\begin{aligned}
(I) &= \left\| F_\varepsilon(S_n + \lambda I)^{\frac{1}{2}-\theta}(f_m - \bar{f}_m) \right\| = \left\| F_\varepsilon(S_n + \lambda I)^{\frac{1}{2}-\theta} q_m(S_n)(S + \lambda I)^{\frac{1}{2}}(S + \lambda I)^{-\frac{1}{2}}(T_n^* \Upsilon - S_n f_{\mathcal{H}}^*) \right\| \\
&\leq \left\| F_\varepsilon(S_n + \lambda I)^{1-\theta} q_m(S_n) \right\| \left\| (S_n + \lambda I)^{-\frac{1}{2}}(S + \lambda I)^{\frac{1}{2}} \right\| \delta(\lambda) \\
&\leq \Lambda \delta(\lambda) \left( \sup_{x \in [0, \varepsilon]} x^{1-\theta} q_m(x) + \lambda^{1-\theta} \sup_{x \in [0, \varepsilon]} q_m(x) \right) \\
&\leq \Lambda \delta(\lambda) \left( \left( \sup_{x \in [0, \varepsilon]} q_m(x) \right)^\theta \left( \sup_{x \in [0, \varepsilon]} x q_m(x) \right)^{1-\theta} + \lambda^{1-\theta} |p'_m(0)| \right) \\
&\leq \Lambda \delta(\lambda) \left( |p'_m(0)|^\theta + \lambda^{1-\theta} |p'_m(0)| \right) \\
&\leq 2\Lambda \delta(\lambda) \tilde{\varepsilon}^{-1} (\lambda + \tilde{\varepsilon})^{1-\theta}.
\end{aligned}$$

The second to last inequality is obtained by the following argument: if  $m \geq 1$ , since  $\varepsilon \leq x_{1,m}$ ,  $p_m$  is decreasing and convex in  $[0, \varepsilon]$  (see Lemma 5.2, point (ii)), we have

$$q_m(x) = \frac{1 - p_m(x)}{x} \leq |p'_m(0)| \quad \text{for } x \in [0, \varepsilon];$$

and also  $x q_m(x) = 1 - p_m(x) \leq 1$  for  $x \in [0, \varepsilon]$ . If  $m = 0$ , we have  $p_0 \equiv 1$  and  $q_m \equiv 0$ , so that  $f_m = \bar{f}_m = 0$  and the above upper bound is also trivially satisfied for any  $\varepsilon > 0$ .

*Second summand:* Using (26), and the fact that  $|p_m(x)| \leq 1$  for  $x \in [0, \varepsilon]$ :

$$\begin{aligned}
(II) &= \left\| F_\varepsilon(S_n + \lambda I)^{\frac{1}{2}-\theta}(\bar{f}_m - f_{\mathcal{H}}^*) \right\| = \left\| F_\varepsilon(S_n + \lambda I)^{\frac{1}{2}-\theta} p_m(S_n) S^\mu w \right\| \\
&\leq \Lambda^2 \left( \varepsilon^\mu (\varepsilon + \lambda)^{\frac{1}{2}-\theta} + c(\mu) Z_\mu(\lambda) (\varepsilon + \lambda)^{\frac{1}{2}-\theta} \right) \|w\| \\
&\leq \Lambda^2 (\varepsilon^\mu + c(\mu) Z_\mu(\lambda)) (\varepsilon + \lambda)^{\frac{1}{2}-\theta} \kappa^{-\mu-\frac{1}{2}} \rho.
\end{aligned}$$

*Third summand:* observe that since  $F_\varepsilon^\perp = \mathbf{1}_{[\varepsilon, \infty)}(S_n)$ , we can write  $F_\varepsilon^\perp = F_\varepsilon^\perp S_n^{-1} S_n$  and

$$\begin{aligned}
(III) &= \left\| F_\varepsilon^\perp(S_n + \lambda I)^{\frac{1}{2}-\theta}(f_m - f_{\mathcal{H}}^*) \right\| \leq \left\| F_\varepsilon^\perp(S_n + \lambda I)^{1-\theta} S_n^{-1} \right\| \left\| F_\varepsilon^\perp(S_n + \lambda I)^{-\frac{1}{2}} S_n(f_m - f_{\mathcal{H}}^*) \right\| \\
&\leq (\varepsilon + \lambda)^{1-\theta} \varepsilon^{-1} \left( \left\| F_\varepsilon^\perp(S_n + \lambda I)^{-\frac{1}{2}} T_n^*(T_n f_m - \Upsilon) \right\| \right. \\
&\quad \left. + \left\| (S_n + \lambda I)^{-\frac{1}{2}}(T_n^* \Upsilon - S_n f_{\mathcal{H}}^*) \right\| \right) \\
&\leq \varepsilon^{-1} (\varepsilon + \lambda)^{\frac{1}{2}-\theta} \|T_n^*(T_n f_m - \Upsilon)\| + \Lambda \varepsilon^{-1} (\varepsilon + \lambda)^{1-\theta} \delta(\lambda).
\end{aligned}$$

Gathering the three terms and rearranging leads to the announced inequality.  $\square$

**Lemma 5.5** (Bounding the residue). *Assume condition  $\mathbf{SC}(r, \rho)$  holds,  $r \geq \frac{1}{2}$ . Let  $\lambda > 0$  be fixed and assume event  $\mathbf{B}(\lambda)$  holds. Then for any iteration step  $1 \leq m \leq n_\Upsilon$ :*

$$\begin{aligned}
\|T_n^*(T_n f_m - \Upsilon)\| &\leq c(\mu) \Lambda^2 \left( |p'_m(0)|^{-(\mu+1)} + Z_\mu(\lambda) |p'_m(0)|^{-1} \right) \kappa^{-\mu-\frac{1}{2}} \rho \\
&\quad + \left( |p'_m(0)|^{-\frac{1}{2}} + \lambda^{\frac{1}{2}} \right) \Lambda \delta(\lambda).
\end{aligned} \tag{27}$$

*Proof.* Using (20) of Lemma 5.2 and the notation therein, it holds

$$\begin{aligned}\|T_n^*(T_n f_m - \Upsilon)\| &= \|p_m(S_n)T_n^*\Upsilon\| \leq \|F_{x_{1,m}}\varphi_m(S_n)T_n^*\Upsilon\| \\ &\leq \|F_{x_{1,m}}\varphi_m(S_n)S_n f_{\mathcal{H}}^*\| + \|F_{x_{1,m}}\varphi_m(S_n)(T_n^*\Upsilon - S_n f_{\mathcal{H}}^*)\| \\ &:= (I) + (II).\end{aligned}$$

We start with controlling the second term:

$$\begin{aligned}(II) &= \|F_{x_{1,m}}\varphi_m(S_n)(T_n^*\Upsilon - S_n f_{\mathcal{H}}^*)\| = \|F_{x_{1,m}}\varphi_m(S_n)(S + \lambda I)^{\frac{1}{2}}(S + \lambda I)^{-\frac{1}{2}}(T_n^*\Upsilon - S_n f_{\mathcal{H}}^*)\| \\ &\leq \|F_{x_{1,m}}\varphi_m(S_n)(S_n + \lambda I)^{\frac{1}{2}}\| \Lambda\delta(\lambda) \\ &\leq \left( \sup_{x \in [0, x_{1,m}]} x^{\frac{1}{2}} \varphi_m(x) + \lambda^{\frac{1}{2}} \sup_{x \in [0, x_{1,m}]} \varphi_m(x) \right) \Lambda\delta(\lambda) \\ &\leq \left( |p'_m(0)|^{-\frac{1}{2}} + \lambda^{\frac{1}{2}} \right) \Lambda\delta(\lambda),\end{aligned}$$

where we used (24), the first component of event  $\mathbf{B}(\lambda)$ , and in the last line inequality (21) with  $\nu = 0, 1$ . For the first term, we use assumption  $\mathbf{SC}(r, \rho)$ , then (26):

$$\begin{aligned}(I) &= \|F_{x_{1,m}}\varphi_m(S_n)S_n f_{\mathcal{H}}^*\| = \|F_{x_{1,m}}\varphi_m(S_n)S_n S^\mu w\| \\ &\leq \Lambda^2 \left( \sup_{t \in [0, x_{1,m}]} t^{\mu+1} \varphi_m(t) + c(\mu)Z_\mu(\lambda) \sup_{t \in [0, x_{1,m}]} t \varphi_m(t) \right) \kappa^{-\mu-\frac{1}{2}}\rho \\ &\leq c(\mu)\Lambda^2 \left( |p'_m(0)|^{-(\mu+1)} + Z_\mu(\lambda) |p'_m(0)|^{-1} \right) \kappa^{-\mu-\frac{1}{2}}\rho,\end{aligned}$$

where for the last inequality we applied (21) with  $\nu = 2(\mu + 1)$ ,  $\nu = 2$ .  $\square$

We now consider the sequence of polynomials  $p_m^{(2)}$  that are orthogonal with respect to the scalar product  $[\cdot, \cdot]_{(2)}$  (see Lemma 5.2, point (iv)). For notational convenience and compatibility below we define  $x_{1,0} = x_{1,0}^{(2)} = \infty$ .

**Lemma 5.6.** *Assume condition  $\mathbf{SC}(r, \rho)$  holds,  $r \geq \frac{1}{2}$ . For any  $\lambda > 0$ , if the event  $\mathbf{B}(\lambda)$  is satisfied, then for any iteration step  $1 \leq m \leq n_\Upsilon$ , and any  $\varepsilon \in (0, x_{1,m-1})$ :*

$$\begin{aligned}[p_{m-1}, p_{m-1}]_{(0)}^{\frac{1}{2}} &= \|p_{m-1}(S_n)T_n^*\Upsilon\| \\ &\leq \Lambda(\varepsilon + \lambda)^{\frac{1}{2}}\delta(\lambda) + c(\mu)\Lambda^2\varepsilon(\varepsilon^\mu + Z_\mu(\lambda))\kappa^{-\mu-\frac{1}{2}}\rho + \varepsilon^{-\frac{1}{2}} \left[ p_{m-1}^{(2)}, p_{m-1}^{(2)} \right]_{(1)}^{\frac{1}{2}}.\end{aligned}\quad (28)$$

*Proof.* By the optimality property defining the CG algorithm,

$$\begin{aligned}\|p_{m-1}(S_n)T_n^*\Upsilon\| &\leq \|p_{m-1}^{(2)}(S_n)T_n^*\Upsilon\| \leq \|F_\varepsilon p_{m-1}^{(2)}(S_n)T_n^*\Upsilon\| + \|F_\varepsilon^\perp p_{m-1}^{(2)}(S_n)T_n^*\Upsilon\| \\ &\leq \|F_\varepsilon T_n^*\Upsilon\| + \varepsilon^{-\frac{1}{2}} \left\| p_{m-1}^{(2)}(S_n)S_n^{\frac{1}{2}}T_n^*\Upsilon \right\| \\ &= \|F_\varepsilon T_n^*\Upsilon\| + \varepsilon^{-\frac{1}{2}} \left[ p_{m-1}^{(2)}, p_{m-1}^{(2)} \right]_{(1)}^{\frac{1}{2}}\end{aligned}$$

For the last inequality, we have used the fact that  $|p_{m-1}^{(2)}(x)| \leq 1$  for  $x \in [0, x_{m-1}^{(2)}]$ , (since  $p_{m-1}^{(2)}(0) = 1$  and  $p_{m-1}^{(2)}$  is nonincreasing on  $[0, x_{m-1}^{(2)}]$ , the case  $m = 1$  being trivial) along with the assumption  $0 < \varepsilon < x_{1,m-1}$ , as well as  $x_{1,m-1} \leq x_{1,m-1}^{(2)}$  (for both of these properties see Lemma 5.2, point (iv)). We now bound

$$\begin{aligned}
\|F_\varepsilon T_n^* \Upsilon\| &\leq \|F_\varepsilon(T_n^* \Upsilon - S_n f_{\mathcal{H}}^*)\| + \|F_\varepsilon S_n f_{\mathcal{H}}^*\| \\
&\leq \left\| F_\varepsilon (S_n + \lambda I)^{\frac{1}{2}} \right\| \left\| (S_n + \lambda I)^{-\frac{1}{2}} (T_n^* \Upsilon - S_n f_{\mathcal{H}}^*) \right\| + \|F_\varepsilon S_n S^\mu w\| \\
&\leq \Lambda(\varepsilon + \lambda)^{\frac{1}{2}} \delta(\lambda) + \|F_\varepsilon S_n S^\mu w\| \\
&\leq \Lambda(\varepsilon + \lambda)^{\frac{1}{2}} \delta(\lambda) + c(\mu) \Lambda^2 \varepsilon (\varepsilon^\mu + Z_\mu(\lambda)) \kappa^{-\mu-\frac{1}{2}} \rho,
\end{aligned}$$

where we have used (26) for the last line.  $\square$

### Proof of Theorem 3.1.

We set

$$\tilde{\lambda}_* = \left( \frac{4D \log 6 \gamma^{-1}}{\sqrt{n}} \right)^{\frac{2}{2\mu+s+1}}, \text{ and } \lambda_* := \kappa \tilde{\lambda}_*. \quad (29)$$

(This normalization was introduced by Caponnetto and Yao, 2010.) The assumed lower bound (10) on  $n$  ensures  $\tilde{\lambda}_* \leq 1$ . We rewrite equivalently the discrepancy stopping rule as follows: for some fixed  $\tau > 0$ ,

$$\hat{m} := \min \left\{ m \geq 0 : \|T_n^*(T_n f_m - \Upsilon)\| \leq (2 + \tau) \lambda_*^{\frac{1}{2}} \delta(\lambda_*) \right\}, \quad (30)$$

where

$$\delta(\lambda_*) := \frac{3}{4} M \tilde{\lambda}_*^{\mu+\frac{1}{2}}. \quad (31)$$

Observe that the above  $\tau > 0$  is related from the constant  $\tau' > 3/2$  in (9) via  $\tau = \frac{4}{3}(\tau' - \frac{3}{2})$ .

We first check that event  $\mathbf{B}(\lambda_*)$ , is satisfied with large probability, using for this concentration results which were recalled in Lemma 5.1. From inequality (17), with probability  $1 - \gamma/3$  we have

$$\begin{aligned}
\left\| (S + \lambda_* I)^{-\frac{1}{2}} (T_n^* \Upsilon - S_n f_{\mathcal{H}}^*) \right\| &\leq 2M \left( \sqrt{\frac{\mathcal{N}(\lambda_*)}{n}} + \frac{2\sqrt{\kappa}}{\sqrt{\lambda_* n}} \right) \log \frac{6}{\gamma} \\
&\leq \frac{2M}{\sqrt{n}} D \tilde{\lambda}_*^{-\frac{s}{2}} \left( 1 + \frac{1}{2D^2} \left( \frac{4D}{\sqrt{n}} \log \frac{6}{\gamma} \right) \tilde{\lambda}_*^{\frac{s-1}{2}} \right) \log \frac{6}{\gamma} \\
&\leq \frac{M}{2} \tilde{\lambda}_*^{\mu+\frac{1}{2}} \left( 1 + \frac{1}{2D^2} \tilde{\lambda}_*^{\mu+s} \right) \\
&\leq \frac{3}{4} M \tilde{\lambda}_*^{\mu+\frac{1}{2}} = \delta(\lambda_*),
\end{aligned} \quad (32)$$

where we have used  $\mathbf{ED}(s, D)$ , (29) and the assumptions  $D \geq 1$  and  $\tilde{\lambda}_* \leq 1$ , as well as the fact that  $\log 6 \gamma^{-1} \geq 1$ . This ensures the first component of  $\mathbf{B}(\lambda_*)$  is satisfied with probability  $1 - \gamma/3$ . We

now turn to the second component. Inequality (18) along with a repetition of the above reasoning yields that with probability  $1 - \gamma/3$ :

$$\left\| (S + \lambda_* I)^{-\frac{1}{2}} (S_n - S) \right\|_{HS} \leq \frac{\sqrt{\kappa}}{M} \delta(\lambda_*),$$

so that

$$\left\| (S + \lambda_* I)^{-1} (S_n - S) \right\| \leq \frac{\sqrt{\kappa}}{M} \lambda_*^{-\frac{1}{2}} \delta(\lambda_*) = \frac{3}{4} \tilde{\lambda}_*^\mu \leq \frac{3}{4}.$$

Observe that

$$(S + \lambda_* I)(S_n + \lambda_* I)^{-1} = ((S_n - S)(S + \lambda_* I)^{-1} + I)^{-1}$$

and use the inequality  $\|(I - A)^{-1}\| = \left\| \sum_{k \geq 0} A^k \right\| \leq (1 - \|A\|)^{-1}$  for  $\|A\| < 1$ , to obtain that the second component in  $\mathbf{B}(\lambda_*)$  is satisfied with  $\Lambda := 2$  (with probability  $1 - \gamma/3$ ).

Finally, equation (19) implies that the third component in  $\mathbf{B}(\lambda_*)$  is satisfied with probability  $1 - \gamma/3$ , with

$$\Delta := 4 \sqrt{\frac{\log 6 \gamma^{-1}}{n}}. \quad (33)$$

To conclude, by the union bound, the three components of event  $\mathbf{B}(\lambda_*)$  are satisfied simultaneously with probability larger than  $1 - \gamma$ . We assume for the rest of the proof that this event is satisfied.

The structure of the proof is now as follows: we aim at bounding the error of the estimator using the inequality of Lemma 5.4. In this upper bound, the residue term is controlled by definition of the stopping rule. The only and most difficult remaining quantity to control is then  $|p'_{\hat{m}}(0)|$ . Using Lemma 5.5 on the residue at iteration  $\hat{m} - 1$ , and the definition of the stopping criterion, will allow to upper bound  $|p'_{\hat{m}-1}(0)|$ ; finally Lemma 5.6 allows to relate iterations  $\hat{m} - 1$  and  $\hat{m}$ .

We will assume  $\hat{m} \geq 1$  for the remainder of the proof and postpone to the end the (simpler) case  $\hat{m} = 0$ .

**First step:** upper bound on  $|p'_{\hat{m}-1}(0)|$ .

By definition of the stopping rule, we have  $\|T_n^*(T_n f_{\hat{m}-1} - \Upsilon)\| > (2 + \tau) \lambda_*^{\frac{1}{2}} \delta(\lambda_*)$ . Now applying this together with the upper bound of Lemma 5.5 and rearranging, we get

$$\begin{aligned} \tau \lambda_*^{\frac{1}{2}} \delta(\lambda_*) &\leq c(\mu) \left( |p'_{\hat{m}-1}(0)|^{-(\mu+1)} + Z_\mu(\lambda_*) |p'_{\hat{m}-1}(0)|^{-1} \right) \kappa^{-\mu-\frac{1}{2}} \rho + 2 |p'_{\hat{m}-1}(0)|^{-\frac{1}{2}} \delta(\lambda_*) \\ &\leq c(\mu) \max \left( |p'_{\hat{m}-1}(0)|^{-\frac{1}{2}} \delta(\lambda_*), \rho \kappa^{-\mu-\frac{1}{2}} |p'_{\hat{m}-1}(0)|^{-(\mu+1)}, \right. \\ &\quad \left. \rho \kappa^{-\mu-\frac{1}{2}} Z_\mu(\lambda_*) |p'_{\hat{m}-1}(0)|^{-1} \right). \end{aligned}$$

We examine in succession the possibilities that the maximum in the above expression is attained for each of the terms which comprise it. If the first term attains the maximum, this implies  $|p'_{\hat{m}-1}(0)| \leq c(\mu) \tau^{-2} \lambda_*^{-1}$ . If the second term attains the maximum, this entails

$$c(\mu) \rho \kappa^{-\mu-\frac{1}{2}} |p'_{\hat{m}-1}(0)|^{-(\mu+1)} \geq \tau \lambda_*^{\frac{1}{2}} \delta(\lambda_*),$$

which using (31) yields  $|p'_{\widehat{m}-1}(0)| \leq c(\mu, \tau) \left(\frac{\rho}{M}\right)^{\frac{1}{\mu+1}} \lambda_*^{-1}$ . Finally, if the third term attains the maximum, we have

$$c(\mu) \rho Z_\mu(\lambda_*) \kappa^{-\mu-\frac{1}{2}} |p'_{\widehat{m}-1}(0)|^{-1} \geq \tau \lambda_*^{\frac{1}{2}} \delta(\lambda_*),$$

which using (31) yields  $|p'_{\widehat{m}-1}(0)| \leq c(\mu, \tau) \frac{\rho}{M} \lambda_*^{-\mu-1} Z_\mu(\lambda_*)$ . We now establish the inequality

$$Z_\mu(\lambda_*) \lambda_*^{-\mu} \leq 1. \quad (34)$$

The inequality is trivial if  $\mu \leq 1$  given the definition of  $Z_\mu(\lambda_*)$  in (23). If  $\mu > 1$  holds, from the definition (33), it holds that  $\Delta \leq \widetilde{\lambda}_*^{\frac{2\mu+s+1}{2}}$  (using  $D \geq 1$ ,  $\log 6\gamma^{-1} \geq 1$ ); hence

$$Z_\mu(\lambda_*) \lambda_*^{-\mu} = \Delta \widetilde{\lambda}_*^{-\mu} \leq \widetilde{\lambda}_*^{\frac{s+1}{2}} \leq 1.$$

Gathering all three cases, we obtain that it always holds that

$$|p'_{\widehat{m}-1}(0)| \leq c(\mu, \tau) \max\left(\frac{\rho}{M}, 1\right) \lambda_*^{-1}. \quad (35)$$

**Second step:** upper bound on  $|p'_{\widehat{m}}(0)|$ . For this we use the result of the first step and relate  $|p'_{\widehat{m}-1}(0)|$  to  $|p'_{\widehat{m}}(0)|$  using property (22) of orthogonal polynomials, which we recall here for convenience:

$$|p_{m-1}'(0) - p_m'(0)| \leq \frac{[p_{m-1}, p_{m-1}]_{(0)}}{[p_{m-1}^{(2)}, p_{m-1}^{(2)}]_{(1)}}. \quad (36)$$

To upper bound the above quantity, we apply Lemma 5.6 with the choice  $\lambda = \lambda_*$  and

$$\varepsilon = \varepsilon_0 := a_0(\mu, \tau) \min\left(\frac{M}{\rho}, 1\right) \lambda_*,$$

where  $0 < a_0(\mu, \tau) \leq 1$  will be chosen small enough in order to satisfy some constraints to be specified below. (We must insist here for the consistency of the argument that contrarily to the notation  $c(\dots)$ , the notation  $a_0(\mu, \tau)$  denotes a fixed value that does not change throughout the proof.) The first constraint is the requirement  $\varepsilon_0 \in (0, x_{1,m-1})$  in order to apply Lemma 5.6. For this, it can be seen from (35) that  $a_0(\mu, \tau)$  can be chosen small enough (namely smaller than the inverse of the constant  $c(\mu, \tau)$  of equation (35)), to ensure

$$\varepsilon_0 \leq |p'_{m-1}(0)|^{-1} \leq x_{1,m-1},$$

the second inequality above is an easy consequence of the fact that  $p_{m-1}$  is convex on  $[0, x_{1,m-1}]$  and  $p_{m-1}(0) = 1$ . We can now apply Lemma 5.6 and use inequality (28). We turn to upper bound the following quantity appearing on the RHS of (28):

$$\begin{aligned} & \Lambda(\varepsilon_0 + \lambda_*)^{\frac{1}{2}} \delta(\lambda_*) + c(\mu) \Lambda^2 \varepsilon_0 (\varepsilon_0^\mu + Z_\mu(\lambda_*)) \kappa^{-\mu-\frac{1}{2}} \rho \\ & \leq 2(a_0(\mu, \tau) + 1)^{\frac{1}{2}} \lambda_*^{\frac{1}{2}} \delta(\lambda_*) + c(\mu) a_0(\mu, \tau) \min(\rho, M) \lambda_*^{\frac{1}{2}} \widetilde{\lambda}_*^{\mu+\frac{1}{2}} \\ & \leq (c(\mu) a_0(\mu, \tau) + 2) \lambda_*^{\frac{1}{2}} \delta(\lambda_*), \end{aligned} \quad (37)$$

where we have used  $\Lambda = 2$ , the definition (31) for  $\delta(\lambda_*)$  and inequality  $Z_\mu(\lambda_*) \leq \lambda_*^\mu$ , see (34). Now, we can choose  $a_0(\mu, \tau)$  small enough so that in addition to the previous constraint, the factor in the last display satisfies  $c(\mu)a_0(\mu, \tau) \leq \frac{\tau}{2}$ . The definition of the stopping rule entails

$$[p_{m-1}, p_{m-1}]_{(0)}^{\frac{1}{2}} = \|T_n^*(T_n f_{\widehat{m}-1} - \Upsilon)\| > (2 + \tau)\lambda_*^{\frac{1}{2}}\delta(\lambda_*). \quad (38)$$

Now combining (28), (38) and (37), we obtain

$$\begin{aligned} [p_{m-1}, p_{m-1}]_{(0)}^{\frac{1}{2}} &\leq (2 + \tau/2)\lambda_*^{\frac{1}{2}}\delta(\lambda_*) + \varepsilon_0^{-\frac{1}{2}} \left[ p_{m-1}^{(2)}, p_{m-1}^{(2)} \right]_{(1)}^{\frac{1}{2}} \\ &\leq \frac{2 + \tau/2}{2 + \tau} [p_{m-1}, p_{m-1}]_{(0)}^{\frac{1}{2}} + \varepsilon_0^{-\frac{1}{2}} \left[ p_{m-1}^{(2)}, p_{m-1}^{(2)} \right]_{(1)}^{\frac{1}{2}}, \end{aligned}$$

so that

$$(2 + 4\tau^{-1})^{-1} [p_{m-1}, p_{m-1}]_{(0)}^{\frac{1}{2}} \leq \varepsilon_0^{-\frac{1}{2}} \left[ p_{m-1}^{(2)}, p_{m-1}^{(2)} \right]_{(1)}^{\frac{1}{2}};$$

using this inequality in relation with (36) and (35), we obtain

$$|p'_{\widehat{m}}(0)| \leq |p'_{\widehat{m}-1}(0)| + c(\tau)\varepsilon_0^{-1} \leq c(\mu, \tau) \max\left(\frac{\rho}{M}, 1\right) \lambda_*^{-1}. \quad (39)$$

**Final step.** We want to apply the main error bound of Lemma 5.4 with  $\lambda = \lambda_*$  and  $\varepsilon = \varepsilon_* = a(\mu, \tau) \min\left(\frac{M}{\rho}, 1\right) \lambda_*$ . Note that  $\varepsilon_*$  is different from  $\varepsilon_0$  considered above; in fact  $\varepsilon_*$  must now satisfy the constraint  $\varepsilon_* \in (0, x_{1,m})$  in order to be able to apply the lemma. In view of (39), we can choose  $a(\mu, \tau) \in (0, 1]$  small enough so as to ensure

$$\varepsilon_* \leq |p'_m(0)|^{-1} \leq x_{1,m},$$

similarly to the previous step (but now at iteration  $m$  instead of  $m-1$ ). Recall that in the notation of Lemma 5.4,  $\widetilde{\varepsilon}_* = \min(\varepsilon_*, |p'_m(0)|^{-1})$ , so that with the above choice we have  $\widetilde{\varepsilon}_* = \varepsilon_*$ . We now apply Lemma 5.4, plug in the inequality (by definition of the stopping rule)

$$\|T_n^*(T_n f_{\widehat{m}} - \Upsilon)\| \leq (2 + \tau)\lambda_*^{\frac{1}{2}}\delta(\lambda_*),$$

and obtain, using again (34):

$$\begin{aligned} \left\| S^{\frac{1}{2}-\theta}(f_m - f_{\mathcal{H}}^*) \right\| &\leq c(\mu, \tau) \left( \widetilde{\varepsilon}_*^{-1} \lambda_*^{1-\theta} \delta(\lambda_*) + \lambda_*^{\frac{1}{2}-\theta+\mu} \kappa^{-\mu-\frac{1}{2}} \rho + \varepsilon_*^{-1} \lambda_*^{1-\theta} \delta(\lambda_*) \right) \\ &\leq c(\mu, \tau) \left( \max\left(\frac{\rho}{M}, 1\right) \lambda_*^{-\theta} \delta(\lambda_*) + \lambda_*^{\frac{1}{2}-\theta+\mu} \kappa^{-\mu-\frac{1}{2}} \rho \right) \\ &\leq c(\mu, \tau) \max(\rho, M) \lambda_*^{-\theta} \widetilde{\lambda}_*^{\mu+\frac{1}{2}} \\ &= c(\mu, \tau) \max(\rho, M) \kappa^{-\theta} \left( \frac{4D}{\sqrt{n}} \log \frac{6}{\gamma} \right)^{\frac{2\mu+1-2\theta}{2\mu+s+1}} \\ &= c(\mu, \tau) \max(\rho, M) \kappa^{-\theta} \left( \frac{4D}{\sqrt{n}} \log \frac{6}{\gamma} \right)^{\frac{2(r-\theta)}{2r+s}}. \end{aligned}$$

If  $\widehat{m} = 0$ , we can apply directly Lemma 5.4 as above without requiring the two previous steps, since in this case  $p'_0(0) = 0$ , so that we obtain the same final bound.

## 5.4 Proof of Theorem 3.2

In the case of the “outer” rates of convergence, i.e. condition  $\mathbf{SC}(r, \rho)$  holds with  $r \in (0, \frac{1}{2})$ , we recall that the target function  $f^*$  is not representable as an element of the Hilbert space  $\mathcal{H}$ . This means many arguments used in Section 5.3 can’t be used directly. To alleviate this, we consider an approximation of  $f^*$  by a function belonging to  $\mathcal{H}$  defined as

$$f_{\mathcal{H}}^{\lambda} := (S_n + \lambda I)^{-1} T^* f^*. \quad (40)$$

Similarly to the previous proof, we define an event where the estimation error is controlled in an appropriate sense:

$$\mathbf{B}'(\lambda) : \begin{cases} \left\| (S + \lambda I)^{-\frac{1}{2}} (T_n^* \Upsilon - T f^*) \right\| & \leq \delta(\lambda), \\ \left\| (S + \lambda I)^{-\frac{1}{2}} (S_n - S) \right\|_{HS} & \leq \tilde{\delta}(\lambda), \\ \left\| (S + \lambda I)(S_n + \lambda I)^{-1} \right\|_{HS} & \leq \Lambda^2, \end{cases}$$

Notice that the first part of the event is slightly different from the corresponding part of  $\mathbf{B}(\lambda)$ ; this is because we will be using concentration inequality (16) rather than (17), the latter only being available for  $r \geq \frac{1}{2}$ .

Our first lemma controls the approximation from  $T f_{\mathcal{H}}^{\lambda}$  to the target  $f^*$ .

**Lemma 5.7.** *Assume condition  $\mathbf{SC}(r, \rho)$  holds,  $r < \frac{1}{2}$ . Let  $\theta$  be fixed,  $\theta \in [0, r)$ . For any  $\lambda > 0$ , if the event  $\mathbf{B}'(\lambda)$  is satisfied, then*

$$\left\| K^{-\theta} (T f_{\mathcal{H}}^{\lambda} - f^*) \right\| \leq \kappa^{-r} \rho \lambda^{r-\theta} \left( 1 + \Lambda^2 \lambda^{-\frac{1}{2}} \tilde{\delta}(\lambda) \right),$$

where  $f_{\mathcal{H}}^{\lambda}$  is defined in (40).

*Proof.* We first write

$$\left\| K^{-\theta} (T f_{\mathcal{H}}^{\lambda} - f^*) \right\| \leq \left\| K^{-\theta} T ((S_n + \lambda)^{-1} - (S + \lambda)^{-1}) T^* f^* \right\| + \left\| K^{-\theta} (T (S + \lambda)^{-1} T^* - I) f^* \right\|$$

We focus on the second term first:

$$\begin{aligned} \left\| K^{-\theta} (T (S + \lambda I)^{-1} T^* - I) f^* \right\| &= \left\| K^{-\theta} (K (K + \lambda I)^{-1} - I) K^r u \right\| \\ &\leq \kappa^{-r} \rho \lambda \sup_{t \in [0, \kappa]} \frac{t^{r-\theta}}{t + \lambda} \\ &\leq \kappa^{-r} \rho \lambda^{r-\theta}, \end{aligned}$$

where at the last line we used Lemma 5.11 and the assumption that  $r \in (0, \frac{1}{2})$  and  $\theta \in (0, r)$  so that  $r - \theta \in (0, \frac{1}{2})$ .



For the first term, we use the second component of  $\mathbf{B}'(\lambda)$ :

$$\begin{aligned}
& \left\| K^{-\theta} T((S_n + \lambda I)^{-1} - (S + \lambda I)^{-1}) T^* f^* \right\| \\
&= \left\| (K^{-\frac{1}{2}} T) S^{\frac{1}{2}-\theta} (S + \lambda I)^{-1} (S - S_n) (S_n + \lambda I)^{-1} T^* K^r u \right\| \\
&\leq \left\| S^{\frac{1}{2}-\theta} (S + \lambda I)^{-\frac{1}{2}} \right\| \left\| (S + \lambda I)^{-\frac{1}{2}} (S - S_n) \right\| \left\| (S_n + \lambda I)^{-1} T^* f^* \right\| \\
&\leq \Lambda^2 \kappa^{-r} \rho \lambda^{r-\theta-\frac{1}{2}} \tilde{\delta}(\lambda);
\end{aligned}$$

for the last inequality, we bounded the last factor by

$$\begin{aligned}
\left\| (S_n + \lambda I)^{-1} T^* f^* \right\| &\leq \left\| (S_n + \lambda I)^{-1} (S + \lambda I) \right\| \left\| (S + \lambda I)^{-1} T^* K^r u \right\| \\
&\leq \Lambda^2 \left\| (S + \lambda I)^{-1} S^{r+\frac{1}{2}} (S^{-\frac{1}{2}} T) u \right\| \\
&\leq \Lambda^2 \kappa^{-r} \rho \sup_{t \in [0, \kappa]} \frac{t^{r+\frac{1}{2}}}{t + \lambda} \\
&\leq \Lambda^2 \kappa^{-r} \rho \lambda^{r-\frac{1}{2}}, \tag{41}
\end{aligned}$$

where we have used Lemma 5.11 again (since  $r + \frac{1}{2} < 1$ ). Collecting the terms yields the conclusion.  $\square$

**Lemma 5.8** (Bounding the error, outer case). *Assume condition  $\mathbf{SC}(r, \rho)$  holds,  $r < \frac{1}{2}$ . For any  $\lambda > 0$ , if the event  $\mathbf{B}'(\lambda)$  is satisfied, then for any  $\theta \in [0, r)$ , for any iteration step  $1 \leq m \leq n_\Upsilon$ , for any  $\varepsilon \in (0, x_{1,m})$ , and denoting  $\tilde{\varepsilon} := \min(\varepsilon, |p'_m(0)|^{-1})$ :*

$$\begin{aligned}
\left\| K^{-\theta} (T f_m - f^*) \right\| &\leq c(\Lambda) \left( \varepsilon^{-1} (\varepsilon + \lambda)^{\frac{1}{2}-\theta} \|T_n^* (T_n f_m - \Upsilon)\| + \tilde{\varepsilon}^{-1} (\lambda + \tilde{\varepsilon})^{1-\theta} \delta(\lambda) \right. \\
&\quad \left. + \kappa^{-r} \rho (\lambda + \varepsilon)^{r-\theta} (1 + \lambda^{-\frac{1}{2}} \tilde{\delta}(\lambda) + \tilde{\varepsilon}^{-1} \lambda) \right)
\end{aligned}$$

For  $m = 0$ , the above inequality is valid for any  $\varepsilon > 0$ .

*Proof.* We begin with

$$\left\| K^{-\theta} (T f_m - f^*) \right\| \leq \left\| K^{-\theta} T (f_m - f_{\mathcal{H}}^\lambda) \right\| + \left\| K^{-\theta} (T f_{\mathcal{H}}^\lambda - f^*) \right\|$$

and the second term is dealt with by Lemma 5.7. For the first term, we will follow the proof of

Lemma 5.4 with appropriate changes. Set  $\tilde{f}_m = q_m(S_n)T^*f^*$ . We have

$$\begin{aligned}
\left\| K^{-\theta} T(f_m - f_{\mathcal{H}}^\lambda) \right\| &= \left\| S^{\frac{1}{2}-\theta}(f_m - f_{\mathcal{H}}^\lambda) \right\| \\
&\leq \Lambda^{1-2\theta} \left\| (S_n + \lambda I)^{\frac{1}{2}-\theta}(f_m - f_{\mathcal{H}}^\lambda) \right\| \\
&\leq \Lambda \left( \left\| F_\varepsilon(S_n + \lambda I)^{\frac{1}{2}-\theta}(f_m - \tilde{f}_m) \right\| + \left\| F_\varepsilon(S_n + \lambda I)^{\frac{1}{2}-\theta}(\tilde{f}_m - f_{\mathcal{H}}^\lambda) \right\| \right. \\
&\quad \left. + \left\| F_\varepsilon^\perp(S_n + \lambda I)^{\frac{1}{2}-\theta}(f_m - f_{\mathcal{H}}^\lambda) \right\| \right) \\
&:= \Lambda((I) + (II) + (III)),
\end{aligned}$$

We upper bound the first summand, using the first component of event  $\mathbf{B}'(\lambda)$ :

$$\begin{aligned}
(I) &= \left\| F_\varepsilon(S_n + \lambda I)^{\frac{1}{2}-\theta} q_m(S_n)(T_n^* \Upsilon - T^* f^*) \right\| \\
&\leq \Lambda \left\| F_\varepsilon(S_n + \lambda I)^{1-\theta} q_m(S_n) \right\| \left\| (S_n + \lambda I)^{-\frac{1}{2}}(T_n^* \Upsilon - T^* f^*) \right\| \\
&\leq 2\Lambda\delta(\lambda) |p'_m(0)| \left( \lambda + |p'_m(0)|^{-1} \right)^{1-\theta};
\end{aligned}$$

the above calculation is almost identical to the handling of term (I) in the proof of Lemma 5.4, and we refer to that proof for the details. We turn to the second term:

$$\begin{aligned}
(II) &= \left\| F_\varepsilon(S_n + \lambda I)^{\frac{1}{2}-\theta}(\tilde{f}_m - f_{\mathcal{H}}^\lambda) \right\| = \left\| F_\varepsilon(S_n + \lambda I)^{\frac{1}{2}-\theta}(q_m(S_n)(S_n + \lambda I) - I)(S_n + \lambda I)^{-1} T^* K^r u \right\| \\
&\leq \left\| F_\varepsilon(S_n + \lambda I)^{-\theta-\frac{1}{2}}(p_m(S_n) + \lambda q_m(S_n)) S^{r+\frac{1}{2}} \right\| \left\| S^{-\frac{1}{2}} T^* u \right\| \\
&\leq \Lambda^2 \kappa^{-r} \rho \sup_{t \in [0, \varepsilon]} (|p_m(t)| + \lambda |q_m(t)|) (t + \lambda)^{r-\theta} \\
&\leq \Lambda^2 \kappa^{-r} \rho \left( (\varepsilon + \lambda)^{r-\theta} + \lambda^{1+r-\theta} |p'_m(0)| + \lambda |p'_m(0)|^{1+\theta-r} \right) \\
&\leq c(\Lambda) \kappa^{-r} \rho (\tilde{\varepsilon} + \lambda)^{r-\theta} (1 + \tilde{\varepsilon}^{-1} \lambda),
\end{aligned}$$

where for the penultimate inequality, we used the same arguments as in the proof of Lemma 5.4 to bound the quantities involving  $|p_m(x)|$  and  $|q_m(x)|$  on the interval  $[0, \varepsilon] \subset [0, x_{1,m}]$ . We finally consider the third term; we recall that we can write  $F_\varepsilon^\perp = F_\varepsilon^\perp S_n^{-1} S_n$  and

$$\begin{aligned}
(III) &= \left\| F_\varepsilon^\perp(S_n + \lambda I)^{\frac{1}{2}-\theta}(f_m - f_{\mathcal{H}}^\lambda) \right\| \leq \left\| F_\varepsilon^\perp(S_n + \lambda I)^{1-\theta} S_n^{-1} \right\| \left\| F_\varepsilon^\perp(S_n + \lambda I)^{-\frac{1}{2}} S_n(f_m - f_{\mathcal{H}}^\lambda) \right\| \\
&\leq \frac{(\varepsilon + \lambda)^{1-\theta}}{\varepsilon} \left( \left\| F_\varepsilon^\perp(S_n + \lambda I)^{-\frac{1}{2}} T_n^*(T_n f_m - \Upsilon) \right\| \right. \\
&\quad \left. + \left\| (S_n + \lambda I)^{-\frac{1}{2}}(T_n^* \Upsilon - T^* f^*) \right\| \right. \\
&\quad \left. + \left\| F_\varepsilon^\perp(S_n + \lambda I)^{-\frac{1}{2}}(T^* f^* - S_n f_{\mathcal{H}}^\lambda) \right\| \right) \\
&\leq \frac{(\varepsilon + \lambda)^{\frac{1}{2}-\theta}}{\varepsilon} \|T_n^*(T_n f_m - \Upsilon)\| + \frac{(\varepsilon + \lambda)^{1-\theta}}{\varepsilon} \delta(\lambda) + (IV),
\end{aligned}$$

with

$$\begin{aligned}
(IV) &:= \varepsilon^{-1}(\varepsilon + \lambda)^{1-\theta} \left\| F_\varepsilon^\perp(S_n + \lambda I)^{-\frac{1}{2}}(T^*f^* - S_n f_{\mathcal{H}}^\lambda) \right\| = \lambda \varepsilon^{-1}(\varepsilon + \lambda)^{1-\theta} \left\| F_\varepsilon^\perp(S_n + \lambda I)^{-\frac{3}{2}} T^*f^* \right\| \\
&\leq \lambda \varepsilon^{-1}(\varepsilon + \lambda)^{\frac{1}{2}-\theta} \left\| (S_n + \lambda I)^{-1} T^*f^* \right\| \\
&\leq \Lambda^2 \kappa^{-r} \rho \varepsilon^{-1}(\varepsilon + \lambda)^{\frac{1}{2}-\theta} \lambda^{r+\frac{1}{2}} \\
&\leq \Lambda^2 \kappa^{-r} \rho (\varepsilon + \lambda)^{r-\theta} (1 + \varepsilon^{-1} \lambda),
\end{aligned}$$

where we have reused inequality (41) at the second to last line. Gathering the different terms now yields the announced inequality.  $\square$

**Lemma 5.9** (Bounding the residue, outer case). *Assume condition  $\mathbf{SC}(r, \rho)$  holds,  $r < \frac{1}{2}$ . Let  $\lambda > 0$  be fixed and assume event  $\mathbf{B}(\lambda)$  holds. Then for any iteration step  $1 \leq m \leq n_\Upsilon$ :*

$$\|T_n^*(T_n f_m - \Upsilon)\| \leq \Lambda^2 \left( 2 |p'_m(0)|^{-(r+\frac{1}{2})} + \lambda^{r+\frac{1}{2}} \right) \kappa^{-r} \rho + \Lambda \left( |p'_m(0)|^{-\frac{1}{2}} + \lambda^{\frac{1}{2}} \right) \delta(\lambda). \quad (42)$$

*Proof.* The proof is similar to that of Lemma 5.5 up to the fact that we use  $T^*f^*$  instead of  $S_n f_{\mathcal{H}}^*$ , so that we skip some details. The main inequality becomes

$$\|T_n^*(T_n f_m - \Upsilon)\| \leq \|F_{x_{1,m}} \varphi_m(S_n) T^*f^*\| + \|F_{x_{1,m}} \varphi_m(S_n) (T_n^* \Upsilon - T^*f^*)\| := (I) + (II),$$

where we used (20) of Lemma 5.2 and the notation therein. The second term is controlled exactly as in the proof of Lemma 5.5, only we use the first component of  $\mathbf{B}'(\lambda)$  instead of that of  $\mathbf{B}(\lambda)$ . It gives rise to

$$(II) \leq \left( |p'_m(0)|^{-\frac{1}{2}} + \lambda^{\frac{1}{2}} \right) \Lambda \delta(\lambda).$$

For the first term, we use assumption  $\mathbf{SC}(r, \rho)$ , then (26) with  $r < \frac{1}{2}$ :

$$\begin{aligned}
(I) &= \|F_{x_{1,m}} \varphi_m(S_n) T^*f^*\| = \|F_{x_{1,m}} \varphi_m(S_n) S^{r+\frac{1}{2}} (S^{-\frac{1}{2}} T^*) u\| \\
&\leq \Lambda^2 \left( \sup_{t \in [0, x_{1,m}]} t^{r+\frac{1}{2}} \varphi_m(t) + \lambda^{r+\frac{1}{2}} \sup_{t \in [0, x_{1,m}]} \varphi_m(t) \right) \kappa^{-r} \rho \\
&\leq \Lambda^2 \left( 2 |p'_m(0)|^{-(r+\frac{1}{2})} + \lambda^{r+\frac{1}{2}} \right) \kappa^{-r} \rho,
\end{aligned}$$

where for the last inequality we applied (21) with  $\nu = 2r + 1 \leq 2$ ,  $\nu = 0$ .  $\square$

Finally, the following lemma is the counterpart of Lemma 5.6 in the outer case:

**Lemma 5.10.** *Assume condition  $\mathbf{SC}(r, \rho)$  holds,  $r < \frac{1}{2}$ . For any  $\lambda > 0$ , if the event  $\mathbf{B}(\lambda)$  is satisfied, then for any iteration step  $1 \leq m \leq n_\Upsilon$ , and any  $\varepsilon \in (0, x_{1,m-1})$ :*

$$\begin{aligned}
[p_{m-1}, p_{m-1}]_{(0)}^{\frac{1}{2}} &= \|p_{m-1}(S_n) T_n^* \Upsilon\| \\
&\leq \Lambda(\varepsilon + \lambda)^{\frac{1}{2}} \delta(\lambda) + \Lambda^2 \left( \varepsilon^{r+\frac{1}{2}} + \lambda^{r+\frac{1}{2}} \right) \kappa^{-r} \rho + \varepsilon^{-\frac{1}{2}} \left[ p_{m-1}^{(2)}, p_{m-1}^{(2)} \right]_{(1)}^{\frac{1}{2}}. \quad (43)
\end{aligned}$$

*Proof.* The first step of the proof is unchanged with respect to that of Lemma 5.6, and we refer to it for the details:

$$\|p_{m-1}(S_n)T_n^*\Upsilon\| \leq \|F_\varepsilon T_n^*\Upsilon\| + \varepsilon^{-\frac{1}{2}} \left[ p_{m-1}^{(2)}, p_{m-1}^{(2)} \right]_{(1)}^{\frac{1}{2}}.$$

Then we follow again the proof of Lemma 5.6, but using  $T^*f^*$  in place of  $S_n f_{\mathcal{H}}^*$ :

$$\begin{aligned} \|F_\varepsilon T_n^*\Upsilon\| &\leq \|F_\varepsilon(T_n^*\Upsilon - T^*f^*)\| + \|F_\varepsilon T^*f^*\| \leq \Lambda(\varepsilon + \lambda)^{\frac{1}{2}}\delta(\lambda) + \left\| F_\varepsilon S^{r+\frac{1}{2}}(S^{-\frac{1}{2}}T^*)u \right\| \\ &\leq \Lambda(\varepsilon + \lambda)^{\frac{1}{2}}\delta(\lambda) + \Lambda^2 \left( \varepsilon^{r+\frac{1}{2}} + \lambda^{r+\frac{1}{2}} \right) \kappa^{-r} \rho, \end{aligned}$$

where we have used (26) (with  $r < \frac{1}{2}$ ) for the last line.  $\square$

We provide for completeness the following simple lemma, which was used a couple of times:

**Lemma 5.11.** *Let  $\lambda > 0$  and  $\nu \in [0, 1]$  be fixed. Then*

$$\sup_{t \in \mathbb{R}_+} \frac{t^\nu}{t + \lambda} = C(\nu)\lambda^{\nu-1},$$

with  $C(\nu) = \nu^\nu(1-\nu)^{(1-\nu)} \in [\frac{1}{2}, 1]$  if  $\nu \in (0, 1)$ , and  $C(0) = 1$ ,  $C(1) = 1$ .

*Proof.* If  $\nu \in (0, 1)$ , the derivative of the function is equal to  $t^{\nu-1}((\nu-1)t + \nu\lambda)/(t + \lambda)^2$ . The value  $t^* := \nu\lambda/(1-\nu)$  is the position of the unique maximum on  $\mathbb{R}_+$ , giving rise to the result. The special cases  $\nu = 0, 1$  are treated easily. Alternatively, the upper bound resulting from  $C(\nu) \leq 1$  can be obtained more directly by using the inequality  $(t + \lambda) \geq t^\nu \lambda^{1-\nu}$ .  $\square$

### Proof of Theorem 3.2

We fix the following values for  $\lambda_*, \tilde{\lambda}_*$  similarly to the inner rate case:

$$\tilde{\lambda}_* = \left( \frac{4D}{\sqrt{n}} \log \frac{4}{\gamma} \right)^{\frac{2}{2r+s}}, \text{ and } \lambda_* := \kappa \tilde{\lambda}_*, \quad (44)$$

satisfying  $\tilde{\lambda}_* \leq 1$  because of assumption (12). The discrepancy stopping rule in the outer case can be rewritten as follows: for some fixed  $\tau > 0$ ,

$$\hat{m} := \min \left\{ m \geq 0 : \|T_n^*(T_n f_m - \Upsilon)\| \leq (8 + \tau) \max \left( 1, \frac{\rho}{M} \right) \lambda_*^{\frac{1}{2}} \delta(\lambda_*) \right\}, \quad (45)$$

where

$$\delta(\lambda_*) := \frac{3}{4} M \tilde{\lambda}_*^r. \quad (46)$$

(Observe that  $\tau' > 6$  from (11) and  $\tau' > 0$  in (45) are related via  $\tau = \frac{4}{3}(\tau' - 6)$ .)

We check that event  $\mathbf{B}'(\lambda_*)$ , is satisfied with large probability. To check the first component, we use (16) instead of (17). Since the easily checked relation  $T_n^* \tilde{\Upsilon} = T_n^* \Upsilon$  holds, we have with probability  $1 - \gamma/2$ ,

$$\left\| (S + \lambda_* I)^{-\frac{1}{2}} (T_n^* \tilde{\Upsilon} - T^* f^*) \right\| = \left\| (S + \lambda_* I)^{-\frac{1}{2}} (T_n^* \Upsilon - T^* f^*) \right\| \leq \frac{3}{4} M \tilde{\lambda}_*^{\mu+\frac{1}{2}} = \delta(\lambda_*), \quad (47)$$

where this inequality follows from identical steps as for (32), to which we refer for details (remember the notation  $\mu = r - \frac{1}{2}$ , there). It is worth noting that in order for the argument leading to (32) to be valid, we need to use the assumption  $r + s \geq \frac{1}{2}$ . This ensures the first component of  $\mathbf{B}'(\lambda_*)$  is satisfied with probability  $1 - \gamma/2$ . We now turn to the second component. We can apply the deviation inequality (18) but with  $n$  replaced by  $\tilde{n}$ , since we make use of all the unlabeled data. Using the fact that  $\frac{n}{\tilde{n}} \leq n^{-\frac{1-2r}{2r+s}} \leq \tilde{\lambda}_*^{1-2r}$ , we obtain that with probability  $1 - \gamma/2$ :

$$\begin{aligned} \left\| (S + \lambda_* I)^{-\frac{1}{2}} (S_{\tilde{n}} - S) \right\|_{HS} &\leq 2\sqrt{\kappa} \left( \sqrt{\frac{N(\lambda_*)}{\tilde{n}}} + \frac{2\sqrt{\kappa}}{\sqrt{\lambda_* \tilde{n}}} \right) \log \frac{4}{\gamma} \\ &\leq \frac{\sqrt{\kappa}}{2} \left( \frac{4D}{\sqrt{n}} \log \frac{4}{\gamma} \right) \left( \tilde{\lambda}_*^{-\frac{s}{2} + \frac{1}{2} - r} + \frac{1}{2D^2} \left( \frac{4D}{\sqrt{n}} \log \frac{4}{\gamma} \right) \tilde{\lambda}_*^{\frac{1}{2} - 2r} \right) \\ &\leq \frac{\sqrt{\kappa}}{2} \tilde{\lambda}_*^{\frac{1}{2}} \left( 1 + \frac{1}{2} \tilde{\lambda}_*^s \right) \\ &\leq \frac{3}{4} \sqrt{\kappa} \tilde{\lambda}_*^{\frac{1}{2}} =: \tilde{\delta}(\lambda_*), \end{aligned}$$

so that the second component of  $\mathbf{B}'(\lambda_*)$  is satisfied with the above value for  $\tilde{\delta}(\lambda)$ ; moreover

$$\left\| (S + \lambda_* I)^{-1} (S_n - S) \right\| \leq \lambda_*^{-\frac{1}{2}} \tilde{\delta}(\lambda_*) = \frac{3}{4},$$

implying (by the same argument as in the proof of Theorem 3.1) that the third component of  $\mathbf{B}'(\lambda_*)$  is satisfied with  $\Lambda := 2$ . We can observe in passing that obtaining the above inequality was the technical reason for introducing the additional unlabeled data in the outer case, since using the labeled data alone would not have granted it for this choice of  $\lambda_*$ . Summarizing, the three components of event  $\mathbf{B}'(\lambda_*)$  are satisfied simultaneously with probability larger than  $1 - \gamma$ . We assume for the rest of the proof that this event is satisfied.

We assume  $\hat{m} \geq 1$  for the remainder of the proof and postpone to the end the (simpler) case  $\hat{m} = 0$ .

**First step:** upper bound on  $|p'_{\hat{m}-1}(0)|$ .

By definition of the stopping rule, we have  $\|T_n^*(T_n f_{\hat{m}-1} - \Upsilon)\| > (8 + \tau) \max(1, \frac{\rho}{M}) \lambda_*^{\frac{1}{2}} \delta(\lambda_*)$ . Applying this together with the upper bound of Lemma 5.9, observing that (44) entails  $\Lambda^2 \lambda_*^{r+\frac{1}{2}} \kappa^{-r} \rho \leq 6\rho \lambda_*^{\frac{1}{2}} \delta(\lambda_*)$ , and rearranging, we get

$$\begin{aligned} \max(1, \frac{\rho}{M}) \tau \lambda_*^{\frac{1}{2}} \delta(\lambda_*) &\leq 8 |p'_{\hat{m}-1}(0)|^{-(r+\frac{1}{2})} \kappa^{-r} \rho + 2 |p'_{\hat{m}-1}(0)|^{-\frac{1}{2}} \delta(\lambda_*) \\ &\leq 10 \max \left( |p'_{\hat{m}-1}(0)|^{-\frac{1}{2}} \delta(\lambda_*), |p'_{\hat{m}-1}(0)|^{-(r+\frac{1}{2})} \kappa^{-r} \rho \right) \end{aligned}$$

If the maximum on the RHS is attained for the first term, this implies

$|p'_{\hat{m}-1}(0)| \leq 4 \min(1, \frac{M}{\rho}) \tau^{-2} \lambda_*^{-1}$ . If the second term attains the maximum, this entails via (46)

$$10\rho \kappa^{-r} |p'_{\hat{m}-1}(0)|^{-(r+\frac{1}{2})} \geq \tau \max(1, \frac{\rho}{M}) \lambda_*^{\frac{1}{2}} \delta(\lambda_*) = \frac{3}{4} \tau \max(M, \rho) \kappa^{-r} \lambda_*^{r+\frac{1}{2}},$$

so that

$$|p'_{\widehat{m}-1}(0)| \leq c(r, \tau) \min(1, \frac{\rho}{M})^{\frac{1}{r+\frac{1}{2}}} \lambda_*^{-1}.$$

Gathering the two cases, we obtain that it always holds that

$$|p'_{\widehat{m}-1}(0)| \leq c(r, \tau) \lambda_*^{-1}. \quad (48)$$

**Second step:** upper bound on  $|p'_{\widehat{m}}(0)|$ . Apply Lemma 5.10 with the choice  $\lambda = \lambda_*$  and

$$\varepsilon = \varepsilon_0 := a_0(r, \tau) \lambda_*,$$

where  $0 < a_0(r, \tau) \leq 1$  will be chosen small enough in order to satisfy some constraints to be specified. The first constraint is the requirement  $\varepsilon_0 \in (0, x_{1,m-1})$  in order to apply Lemma 5.10. For this, it can be seen from (48) that  $a_0(r, \tau)$  can be chosen small enough to ensure

$$\varepsilon_0 \leq |p'_{m-1}(0)|^{-1} \leq x_{1,m-1}.$$

We can now apply Lemma 5.10. We upper bound the following quantity appearing on the RHS of (43):

$$\begin{aligned} \Lambda(\varepsilon_0 + \lambda_*)^{\frac{1}{2}} \delta(\lambda_*) + \Lambda^2 \left( \varepsilon_0^{r+\frac{1}{2}} + \lambda_*^{r+\frac{1}{2}} \right) \kappa^{-r} \rho &\leq 2(a_0(r, \tau) + 1) \lambda_*^{\frac{1}{2}} \delta(\lambda_*) + 4(a_0(r, \tau)^{r+\frac{1}{2}} + 1) \lambda_*^{\frac{1}{2}} \widetilde{\lambda}_*^r \rho \\ &\leq (8 + 8a_0(r, \tau)^{\frac{1}{2}}) \max(1, \frac{\rho}{M}) \lambda_*^{\frac{1}{2}} \delta(\lambda_*), \end{aligned} \quad (49)$$

We can choose  $a_0(r, \tau)$  small enough so that in addition to the previous constraint, it satisfies  $a_0(\mu, \tau)^{\frac{1}{2}} \leq \frac{\tau}{16}$ . Remember that the definition of the stopping rule entails

$$[p_{m-1}, p_{m-1}]_{(0)}^{\frac{1}{2}} = \|T_n^*(T_n f_{\widehat{m}-1} - \Upsilon)\| > (8 + \tau) \max(1, \frac{\rho}{M}) \lambda_*^{\frac{1}{2}} \delta(\lambda_*). \quad (50)$$

Combining (43), (50) and (49), we obtain

$$\begin{aligned} [p_{m-1}, p_{m-1}]_{(0)}^{\frac{1}{2}} &\leq (8 + \tau/2) \max(1, \frac{\rho}{M}) \lambda_*^{\frac{1}{2}} \delta(\lambda_*) + \varepsilon_0^{-\frac{1}{2}} \left[ p_{m-1}^{(2)}, p_{m-1}^{(2)} \right]_{(1)}^{\frac{1}{2}} \\ &\leq \frac{8 + \tau/2}{8 + \tau} [p_{m-1}, p_{m-1}]_{(0)}^{\frac{1}{2}} + \varepsilon_0^{-\frac{1}{2}} \left[ p_{m-1}^{(2)}, p_{m-1}^{(2)} \right]_{(1)}^{\frac{1}{2}} \end{aligned}$$

so that

$$(2 + 16\tau^{-1})^{-1} [p_{m-1}, p_{m-1}]_{(0)}^{\frac{1}{2}} \leq \varepsilon_0^{-\frac{1}{2}} \left[ p_{m-1}^{(2)}, p_{m-1}^{(2)} \right]_{(1)}^{\frac{1}{2}};$$

using this inequality in relation with (22) and (48), we obtain

$$|p'_{\widehat{m}}(0)| \leq |p'_{\widehat{m}-1}(0)| + c(\tau) \varepsilon_0^{-1} \leq c(r, \tau) \lambda_*^{-1}. \quad (51)$$

**Final step.** We want to apply the main error bound of Lemma 5.4 with  $\lambda = \lambda_*$  and  $\varepsilon = \varepsilon_* = a(r, \tau) \lambda_*$ . In view of (51), we can choose  $a(\mu, \tau) \in (0, 1]$  small enough so that to ensure

$$\varepsilon_* \leq |p'_m(0)|^{-1} \leq x_{1,m}.$$

Recall that in the notation of Lemma 5.8,  $\tilde{\varepsilon}_* = \min(\varepsilon_*, |p'_m(0)|^{-1})$ , so that with the above choice we have  $\tilde{\varepsilon}_* = \varepsilon_*$ . We now apply Lemma 5.8, plug in the inequality (by definition of the stopping rule)

$$\|T_n^*(T_n f_{\hat{m}} - \Upsilon)\| \leq (8 + \tau) \max(1, \frac{\rho}{M}) \lambda_*^{\frac{1}{2}} \delta(\lambda_*),$$

to obtain:

$$\begin{aligned} \|K^{-\theta}(f_{\hat{m}} - f^*)\| &\leq c(r, \tau) \left( \varepsilon_*^{-1} (\varepsilon_* + \lambda_*)^{\frac{1}{2}-\theta} \max(1, \frac{\rho}{M}) \lambda_*^{\frac{1}{2}} \delta(\lambda_*) + \tilde{\varepsilon}_*^{-1} (\lambda_* + \tilde{\varepsilon}_*)^{1-\theta} \delta(\lambda_*) \right. \\ &\quad \left. + \kappa^{-r} \rho (\lambda_* + \varepsilon_*)^{r-\theta} (1 + \lambda_*^{-\frac{1}{2}} \tilde{\delta}(\lambda) + \tilde{\varepsilon}_*^{-1} \lambda_*) \right) \\ &\leq c(r, \tau) \max(\rho, M) \kappa^{-\theta} \lambda_*^{r-\theta} \\ &= c(r, \tau) \max(\rho, M) \kappa^{-\theta} \left( \frac{4D}{\sqrt{n}} \log \frac{4}{\gamma} \right)^{\frac{2(r-\theta)}{2r+s}}. \end{aligned}$$

If  $\hat{m} = 0$ , we can apply directly Lemma 5.8 as above without requiring the two previous steps, since in this case  $p'_0(0) = 0$ , so that we obtain the same final bound.

## References

- R. Bathia. *Matrix Analysis*, volume 169 of *Graduate texts in mathematics*. Springer, 1997.
- F. Bauer, S. Pereverzev, and L. Rosasco. On Regularization Algorithms in Learning Theory. *Journal of Complexity*, 23:52–72, 2007.
- G. Blanchard and N. Krämer. Optimal learning rates for kernel conjugate gradient regression. In *Advances in Neural Inf. Proc. Systems (NIPS 2010)*, pages 226–234, 2011.
- G. Blanchard and P. Massart. Discussion of "2004 IMS medallion lecture: Local Rademacher complexities and oracle inequalities in risk minimization", by V. Koltchinskii. *Annals of Statistics*, 34(6):2664–2671, 2006.
- G. Blanchard and N. Mücke. Optimal rates for regularization of statistical inverse learning problems. Technical report, University of Potsdam, 2016. (arXiv:1604.04054).
- A. Caponnetto and E. De Vito. Optimal Rates for Regularized Least-squares Algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.
- A. Caponnetto and Y. Yao. Cross-validation based Adaptation for Regularization Operators in Learning Theory. *Analysis and Applications*, 8(2):161–183, 2010.
- N. Cristianini and J. Shawe-Taylor. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.



- E. De Vito, L. Rosasco, A. Caponnetto, U. De Giovannini, and F. Odone. Learning from Examples as an Inverse Problem. *Journal of Machine Learning Research*, 6(1):883, 2006.
- L. Dicker, D. Foster, and D. Hsu. Kernel methods and regularization techniques for nonparametric regression: Minimax optimality and adaptation. Technical report, Rutgers University, 2015. <http://www.stat.rutgers.edu/home/ldicker/papers/kernels.pdf>.
- H.W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Kluwer Academic Publishers, 1996.
- M. Hanke. *Conjugate Gradient Type Methods for Linear Ill-posed Problems*. Pitman Research Notes in Mathematics Series, 327, 1995.
- L. Lo Gerfo, L. Rosasco, E. Odone, F. and De Vito, and A. Verri. Spectral Algorithms for Supervised Learning. *Neural Computation*, 20:1873–1897, 2008.
- A. S. Nemirovskii. The Regularizing Properties of the Adjoint Gradient Method in Ill-posed Problems. *USSR Computational Mathematics and Mathematical Physics*, 26(2):7–16, 1986.
- I. F. Pinelis and A. I. Sakhanenko. Remarks on inequalities for probabilities of large deviations. *Theory Probab. Appl.*, 1(30):143–148, 1985.
- R. Rosipal and L.J. Trejo. Kernel Partial Least Squares Regression in Reproducing Kernel Hilbert Spaces. *Journal of Machine Learning Research*, 2:97–123, 2001.
- S. Smale and D. Zhou. Learning theory estimates via integral operators and their approximation. *Constructive Approximation*, 26(2):153–172, 2007.
- I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, 2008.
- S. Wold, H. Ruhe, H. Wold, and W.J. Dunn III. The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses. *SIAM Journal of Scientific and Statistical Computations*, 5:735–743, 1984.
- Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.
- V. Yurinski. *Sums and Gaussian vectors*, volume 1617 of *Lecture notes in mathematics*. Springer, 1995.
- T. Zhang. Learning bounds for kernel regression using effective data dimensionality. *Neural Computation*, 17(9):2077–2098, 2005.